Technische Universität Ilmenau
Fakultät für Mathematik
und Naturwissenschaften
Institut für Mathematik
http://www.mathematik.tu-ilmenau.de/Math-Net/index_de.html

Postfach 10 05 65
D - 98684 Ilmenau
Germany
Tel.: 03677/69 3267
Fax: 03677/69 3272
Telex: 33 84 23 tuil d.
email: werner.neundorf@tu-ilmenau.de

Preprint No. M 04/04

Grundlagen der numerischen linearen Algebra

Werner Neundorf

Februar 2004

‡MSC (2000): 08-01, 15-01, 65-01, 65F05, 65F15, 65F35

Zusammenfassung

Die Arbeit enthält zahlreiche Informationen zu Grundlagen der Numerischen Linearen Algebra. Als Zusammenstellung von Basiswissen kann es dem Leser für die weitere Beschäftigung mit unterschiedlichen Fragestellungen aus diesem Bereich dienen, insbesondere zu solchen Kapiteln wie

- Generierung von Feldern,
- sparse Matrizen,
- Multiplikation mit Matrizen,
- Matrixskalierung,
- Matrixtransformation und -faktorisierung,
- direkte Verfahren für lineare Gleichungssysteme,
- iterative Verfahren für lineare Gleichungssysteme,
- Abstiegsverfahren für lineare Gleichungssysteme,
- Bandbreitenreduktion von Matrizen.

Zu Beginn des Preprints werden einige Beispiele für das Auftreten linearer Gleichungssysteme erläutert.

Den Schwerpunkt bilden jedoch Aussagen zu Vektoren, Matrizen und Gleichungssystemen, die zum großen Teil als Sätze formuliert und bewiesen sind. Für eine Reihe der vorgestellten Eigenschaften, Methoden und Verfahren sind die Vorgehensweisen durch Schemata, Bilder oder Beispiele veranschaulicht. Dies soll das Verstehen der Zusammenhänge vertiefen und erleichtern. Die Darstellung ist systematisch, verständlich und sehr anschaulich. Vorausgesetzt werden nur wenige Grundkenntnisse der Analysis und linearen Algebra.

Die behandelten computer- und sprachspezifischen Aspekte sowie numerischen Methoden werden jeweils anhand einzelner Beispiele erklärt und zum Teil mit den Computeralgebrasystemen Maple, Matlab und/oder in einer höheren Programmiersprache durchgerechnet. Algorithmen sind teilweise auch im Pseudokode notiert.

Die behandelten Themen eigenen sich sowohl für Studenten des Diplomstudienganges Mathematik im Grund- und Hauptstudium als auch für Studenten der Informatik, Ingenieurwissenschaften und der Wirtschaftswissenschaften im Hauptstudium, die zum Beispiel den Kurs “Numerik großer Gleichungsysteme“ belegen, wie er im Nebenfach Mathematik für Studenten der Informatik und Ingenieurinformatik an der TU Ilmenau angeboten wird.

Ich möchte noch erwähnen, dass dieses Preprint ein guter Einstieg ist für die Numerik von Matrizen und Gleichungssystemen, wie sie in den Lehrbüchern von M. Hermann *Numerische Mathematik*, A. Meister *Numerik linearer Gleichungssysteme*, W. Neundorf *Numerische Mathematik* und R. Plato *Numerische Mathematik kompakt* dargestellt ist.

Inhaltsverzeichnis

Kapitel 1

Beispiele für das Auftreten linearer Gleichungssysteme

1.1 Randwertaufgabe – Stabdurchbiegung

Als Problem wird eine einfache eindimensionale Zweipunktrandwertaufgabe mit inhomogenen Randbedingungen gewählt. Sie ist in sogenannter Divergenzform gegeben, der Differentialoperator als gewöhnliche zweite Ableitung ist selbstadjungiert.
Das zur numerischen Behandlung der Differentialgleichung (DGL) verwendete Diskretisierungsverfahren ist die finite Differenzenmethode (FDM) [39]. Diese führt auf ein lineares Gleichungssystem (LGS), dessen Eigenschaften erläutert werden. Des Weiteren wird für die Lösung des LGS ein Iterationsverfahren (IV) untersucht, dabei werden insbesondere die Fragen der Konvergenz, im Zusammenhang mit dem Spektrum der jeweiligen Iterationsmatrix, sowie die Effizienz des Verfahrens erläutert.

- Zweipunktrandwertaufgabe (RWA) mit inhomogenen Randbedingungen:

$$\begin{aligned} -U''(x) &= F(x), \quad x \in \Omega = (0,1) \subset \mathbb{R}, \\ U(x) &= \varphi(x) \text{ für } x \in \partial\Omega \quad \text{bzw. } U(0) = \varphi_0, \quad U(1) = \varphi_1. \end{aligned} \tag{1.1}$$

- Gitter: $\overline{\Omega}_h = \{x \mid x = x_i = ih,\ i = 0(1)N,\ h = 1/N\}$, h Maschenweite.
- Gitterfunktion: $u_h = (u_1, u_2, ..., u_{N-1})^T$ mit $u_i \approx U_i = U(ih)$.
- Analog für rechte Seite: $f_h = (f_1, f_2, ..., f_{N-1})^T$ mit $f_i = F_i = F(ih)$, d. h. auf dem Gitter wird die rechte Seite exakt dargestellt.
- Approximation der Ableitungen (Operatoren) mittels Differenzenausdrücken:

  $U''(x_i) \approx \frac{1}{h^2}(U_{i+1} - 2U_i + U_{i-1})$ zentraler Differenzenquotient 2. Ordnung.

  Damit haben wir die Konsistenz bzw. Genauigkeit mit der Ordnung $\mathcal{O}(h^2)$.

- Diskretisierte Aufgabe als LGS:

$$\begin{aligned} -\frac{1}{h^2}(u_{i+1} - 2u_i + u_{i-1}) &= f_i, \quad i = 1, 2, ..., N-1, \\ u_0 = \varphi_0, \quad u_N &= \varphi_1. \end{aligned} \tag{1.2}$$

- Matrixschreibweise des LGS:

$$A_h u_h = b_h \quad \text{bzw.} \quad Au = b \tag{1.3}$$

mit

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}, \quad b_h = \begin{pmatrix} f_1 + \varphi_0/h^2 \\ f_2 \\ f_3 \\ \cdots \\ f_{N-2} \\ f_{N-1} + \varphi_1/h^2 \end{pmatrix},$$

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}, \quad b = h^2 \begin{pmatrix} f_1 + \varphi_0/h^2 \\ f_2 \\ f_3 \\ \cdots \\ f_{N-2} \\ f_{N-1} + \varphi_1/h^2 \end{pmatrix},$$

$$A = \text{tridiag}(-1, 2, -1).$$

Die Koeffizientenmatrix $A(n, n) = (a_{ij})$ $(n = N - 1)$ ist symmetrisch sowie positiv definit (spd) und damit regulär.

Sie ist schwach besetzt (sparse Matrix) und hat etwa $3n$ nicht verschwindende Elemente (Nichtnullelemente, NNE) anstelle von n^2 Elementen bei voll besetzten Matrizen. Sie ist eine Tridiagonalmatrix, d. h. ihre Bandbreite ist 3.

Wir erkennen eine besondere Vorzeichensituation der Matrixelemente a_{ij} und für die voll besetzte inverse Matrix $A^{-1} = (a'_{ij})$ gilt $a'_{ij} \geq 0$, was mit weiteren Matrixeigenschaften, wie L-Matrix, M-Matrix, Monotonie u. a. zusammenhängt.

Die Eigenwerte (EW) der Matrix A sind

$$\lambda_i = 2[1 - \cos(i\pi/(n+1))] = 4\sin^2(i\pi/(2(n+1))), \quad i = 1, 2, ..., n,$$

die Eigenvektoren (EV)

$$v^{(i)} = (v_1^{(i)}, v_2^{(i)}, ..., v_n^{(i)})^T \quad \text{mit} \quad v_j^{(i)} = \sin(ij\pi/(n+1)).$$

Wegen $h = 1/(n+1)$, $\cos(x) = 1 - 2\sin^2(x/2)$, gelten die Beziehungen

$$\lambda_i = 4\sin^2(i\pi h/2),$$

$$v^{(i)}, \quad v_j^{(i)} = \sin(ij\pi h),$$

$$0 < 2(1-\cos(\pi h)) = 4\sin^2(\pi h/2) = \lambda_{min} = \lambda_1 < \lambda_2 < ... < \lambda_n = \lambda_{max} =$$
$$= 4\sin^2(n\pi h/2) = 4[1-\sin^2(\pi h/2)] = 4\cos^2(\pi h/2) = 4 - \lambda_{min} < 4,$$

$$\lambda_1 \approx \pi^2 h^2, \qquad \lim_{h\to 0} \lambda_1 = 0,$$

$$\lambda_n \approx 4 - \pi^2 h^2, \quad \lim_{h\to 0} \lambda_n = 4,$$

$$\lambda_{\frac{n+1}{2}} = 2, \quad \text{falls } n \text{ ungerade.}$$

Damit ist das Spektrum $\sigma(A) \in (0,4)$ und beschränkt, was auch die gleichmäßige Beschränkheit der Matrix A für hinreichend kleines h, also $h \le h_0$, bedeutet. Solche Eigenschaften werden im Zusammenhang mit der Untersuchung der Stabilität eines Verfahrens verwendet.

Das IV für die Lösung von $Au = b$ notieren wir in der Basisversion gemäß [13] als

$$u^{(m+1)} = Hu^{(m)} + c = (I - W^{-1}A)u^{(m)} + W^{-1}b = u^{(m)} + W^{-1}r^{(m)},\ m = 0, 1, ..., \quad (1.4)$$

wobei

$u^{(0)}$	Startvektor,
$r^{(m)} = b - Au^{(m)}$	Residuum (manchmal auch $r^{(m)} = Au^{(m)} - b$),
$Au^{(m)} - b$	Defekt,
W	Wichtung, Vorkonditionierungsmatrix, Präkonditionierer,
$W^{-1}(Au^{(m)} - b)$	Korrekturvektor,
$H = I - W^{-1}A$	Iterationsmatrix bedeuten.

Konvergenz des IV

Konvergenzsätze liefern hinreichende und notwendige Konvergenzbedingungen für das IV. Dazu benötigen wir die EW der Iterationsmatrix H oder Näherungen dieser. Hier ist die Konvergenz im Fall des Gesamtschrittverfahrens (Jacobi-Verfahren, GSV) mit $W = D = \text{diag}(A)$ gesichert durch die Aussage, dass die Koeffizientenmatrix A eine irreduzibel diagonaldominante Matrix darstellt.

Die reelle Matrix $A = A(n,n)$ heißt irreduzibel diagonaldominant, wenn

$$(1)\ \ |a_{ii}| \ \ge \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, 2, ..., n,$$

und für mindestens ein i gilt die strenge Größerbeziehung,

(2) und es gibt keine Permutationsmatrix P, mit der eine Transformation der Matrix gemäß

$$\widetilde{A} \ = \ PAP^T \ = \ \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{pmatrix}, \quad \widetilde{A}_{11} \in \mathbb{R}^{p,p},\ \widetilde{A}_{22} \in \mathbb{R}^{q,q},\ p+q = n,$$

möglich ist (falls es eine solche Transformation gibt, ist die Matrix reduzibel).

Dann gelten für A auch die folgenden Aussagen:

(1) $a_{ii} \neq 0$ für alle $i = 1, 2, ..., n$,

(2) Determinante $\det(A) \neq 0$.

Die Iterationsmatrix $H = J = I - D^{-1}A = I - \frac{1}{2}A$ hat die EW

$$\mu_i = \mu_i(H) = 1 - \frac{1}{2}\lambda_i \in (-1, 1),\ \mu_i = -\mu_{n-i},$$

wobei λ_i die EW von A sind. Es gilt wegen

$$\begin{aligned} \lambda_1 &= \lambda_{min} = 4\sin^2(\pi h/2), \\ \lambda_n &= \lambda_{max} = 4 - \lambda_{min} = 4\cos^2(\pi h/2) \end{aligned}$$

für den Spektralradius

$$\rho(H) = \max|\mu(H)| = \mu_1 = 1 - \frac{1}{2}\lambda_{min} = \left|1 - \frac{1}{2}\lambda_{max}\right| = 1 - 2\sin^2\left(\frac{\pi h}{2}\right) = \cos(\pi h),$$

also näherungsweise $\rho(H) \approx 1 - \pi^2 h^2/2 < 1$.

Der Spektralradius liegt aber für kleine Schrittweiten h nahe der Eins, so dass die Konvergenzgeschwindigkeit bzw. Konvergenzrate des IV klein ist.

Als direktes Verfahren zur Lösung von $Au = b$ kann man die Gauß-Elimination mit gleichzeitiger LU-Faktorisierung verwenden, die wegen der speziellen Matrixform in verkürzter vektorisierter Variante ohne Pivotstrategie durchführbar ist. Leider ist die Kondition der Matrix A sehr schlecht, d. h. wir haben mit der Spektralnorm

$$\begin{aligned} \|A\|_2 &= \sqrt{\max_{i=1(1)n} \mu_i},\quad 0 \leq \mu_i \in \sigma(A^T A), \\ &= \sqrt{\rho(A^T A)}, \\ &\qquad \sigma(A^T A) = \{\mu_i(A^T A),\ i = 1, 2, ..., n\} \ \text{ Spektrum}, \\ &\qquad \rho(A^T A) = \max_{i=1(1)n} |\mu_i(A^T A)| \ \text{ Spektralradius}, \end{aligned}$$

wegen $A = A^T > 0$ die spektrale Kondition

$$\begin{aligned} \kappa(A) &= \text{cond}_2(A) = \|A\|_2 \, \|A^{-1}\|_2 \\ &= \frac{\lambda_{max}}{\lambda_{min}} = \frac{4 - \lambda_{min}}{\lambda_{min}} = \frac{4}{\lambda_{min}} - 1 \\ &\approx \frac{4}{4\sin^2(\pi h/2)} \approx \frac{1}{(\pi h/2)^2} \\ &\approx \frac{4n^2}{\pi^2} \gg 1 \ \text{ für } \ n \gg 1. \end{aligned}$$

Um eine akzeptable Näherungslösung u zur exakten u^* zu erhalten, ist eine starke Gleitpunktarithmetik (GPA) notwendig.
Die Fehlerakkumulation beim Gauß-Algorithmus (Gauß-Elimination, GA) mit $A(n,n)$ und t Dualstellen der Mantisse des Gleitpunktformats (GPF) führt bei $n \to \infty$ zum absoluten Fehler

$$\|\delta u\| = \|u^* - u\| = 2^{-t}\text{cond}(A)\ K(n),$$

wobei

$$K(n) = \begin{cases} \mathcal{O}(\sqrt{n}) & \text{für verkürzten GA ohne Pivotisierung bei } A = \text{tridiag}(), \text{ dazu} \\ & A \text{ diagonaldominant oder } A = A^T > 0, \\ \mathcal{O}(n) & \text{für GA ohne Pivotisierung, } A \text{ diag.dominant oder } A = A^T > 0, \\ \mathcal{O}(2^n) & \text{für GA mit Spaltenpivotisierung,} \\ \mathcal{O}(n^{3/2}) & \text{für GA mit vollständiger Pivotisierung.} \end{cases}$$

Für $t = 64$ (*extended*-Format), $t = 53$ (*double*-Format) oder $t = 40$ (*real*-Format) kann man die gültigen Dezimalstellen der Näherungslösung ermitteln.

1.2 Diffusion-Konvektions-Randwertaufgabe

Die Modellprobleme sind spezielle singulär gestörte RWA mit inhomogenen Randbedingungen auf verschiedenen Bereichen. Die Störung wird durch einen kleinen reellen Parameter $\varepsilon,\ 0 < \varepsilon \ll 1$, im Diffusionsanteil der Gleichung charakterisiert. Der Konvektionsterm hängt mit dem Auftreten der ersten Ableitung zusammen, so dass damit keine Divergenzform der DGL vorliegt.
Die numerische Lösung erfolgt mit verschiedenen Diskretisierungsverfahren.

- Singulär gestörte RWA mit inhomogenen Randbedingungen:

$$\begin{aligned} -\varepsilon U''(x) - 2U'(x) &= 0, \quad x \in \Omega = (0,b) \subset \mathbb{R}, \\ U(x) &= \varphi(x) \ \text{ für } x \in \partial\Omega \ \text{ bzw. } U(0)=\varphi_0,\ U(b)=\varphi_1. \end{aligned} \tag{1.5}$$

  Für konkrete Randbedingungen kann man die exakte Lösung der DGL angeben, was in Testrechnungen zur Überprüfung der Genauigkeit des Näherungsverfahrens (Konsistenzordnung) und damit der Näherungslösung ausgenutzt werden kann.
  So ergeben

$$b = \infty,\ U(0) = 1,\ U(\infty) = 0 \quad \Rightarrow \quad U(x) = e^{-2x/\varepsilon},$$

$$b = 1, \quad U(0) = 1, \quad U(1) = 0 \quad \Rightarrow \quad U(x) = \frac{e^{-2x/\varepsilon} - e^{-2/\varepsilon}}{1 - e^{-2/\varepsilon}}.$$

  Wir werden zunächst jeweils beide Intervalle mit den Lösungen betrachten.

(A) FDM

- Gitter auf $[0,\infty]$: $\overline{\Omega}_h = \{x \mid x = x_i = ih,\ i = 0,1,2,....,\ h > 0\}$,
  Gitter auf $[0,1]$: $\overline{\Omega}_h = \{x \mid x = x_i = ih,\ i = 0(1)N,\ h = 1/N\}$.
- Gitterfunktion: $u_h = (u_1, u_2, ..., u_{N(\infty)-1})^T$ mit $u_i \approx U_i = U(ih)$.
- Da es sich um eine homogene RWA handelt, kann die rechte Seite auf dem Gitter exakt dargestellt werden.
- Die Approximation der Ableitungen (Operatoren) mittels Differenzenausdrücken berücksichtigt für die erste Ableitung mehrere Möglichkeiten:

$$(1) \quad U''(x_i) \approx \frac{1}{h^2}(U_{i+1} - 2U_i + U_{i-1}) \quad \text{zentraler Differenzenquotient}$$

mit Genauigkeitsordnung $\mathcal{O}(h^2)$,

$$(2) \quad U'(x_i) \approx \frac{1}{2h}(U_{i+1} - U_{i-1}) \quad \text{zentraler Differenzenquotient}$$

mit Genauigkeitsordnung $\mathcal{O}(h^2)$,

$$(3) \quad U'(x_i) \approx \frac{1}{h}(U_{i+1} - U_i) \quad \text{Vorwärtsdifferenzenquotient}$$

mit Genauigkeitsordnung $\mathcal{O}(h)$,

$$(4) \quad U'(x_i) \approx \frac{1}{h}(U_i - U_{i-1}) \quad \text{Rückwärtsdifferenzenquotient}$$

mit Genauigkeitsordnung $\mathcal{O}(h)$.
Man kann vermuten, dass für akzeptable Ergebnisse der Näherung auf dem Gitter die Maschenweite h hinreichend klein und auch kleiner als der Parameter ε zu wählen ist.

- So entstehen Differenzenschemata (DS) für die verschiedenen Kombinationen. Die Differenzengleichungen, multipliziert mit h^2, sind homogen und haben konstante Koeffizienten. So ist eine exakte Lösung möglich. Mit speziellen Ansätzen der Form $u_i = c\rho^i$ lassen sich diese in quadratische Gleichungen bezüglich ρ überführen. In der allgemeinen Lösung $u_i = c_1\rho_1^i + c_2\rho_2^i$ ergeben sich dann noch die Größen c_i aus den zu erfüllenden Randbedingungen.
  Variante (1)+(2): DS mit zentralen Differenzenquotienten

$$\begin{aligned} -(\varepsilon + h)u_{i+1} + 2\varepsilon u_i - (\varepsilon - h)u_{i-1} &= 0, \quad i = 1,2,...,N(\infty) - 1, \\ u_0 = 1,\ u_\infty = 0 \qquad &\text{bzw.} \quad u_0 = 1,\ u_N = 0. \end{aligned} \tag{1.6}$$

Die exakten Lösungen der DS sind

$$u_i = \Big(\frac{\varepsilon - h}{\varepsilon + h}\Big)^i, \quad i = 0, 1, ..., \infty,$$

$$u_i = \frac{\mu^i - \mu^N}{1 - \mu^N}, \quad \mu = \frac{\varepsilon - h}{\varepsilon + h}, \quad i = 0, 1, ..., N.$$

Damit keine oszillierenden (parasitären) Anteile in der Lösung auftreten, was ja von der Physik her keinen Sinn machen würde, muss $h < \varepsilon$ sein. Damit ist auch $\mu < 1$. Diese Bedingung widerspiegelt sich sowohl in der Ungleichung $R_h = \frac{h}{\varepsilon} \leq 1$ für die diskrete Reynolds-Zahl R_h (physikalische Kennzahl) als auch bei der Betrachtung der Fehlerordnung $\mathcal{O}((\frac{h}{\varepsilon})^2)$ der Lösung u_i.

Variante (1)+(3): Upwind-DS bei Beachtung des Vorzeichens vor U'

$$\begin{aligned} -(\varepsilon + 2h)u_{i+1} + (2\varepsilon + 2h)u_i - \varepsilon u_{i-1} &= 0, \quad i = 1, 2, ..., N(\infty) - 1, \\ u_0 = 1, \ u_\infty = 0 \qquad &\text{bzw.} \quad u_0 = 1, \ u_N = 0. \end{aligned} \tag{1.7}$$

Die exakten Lösungen der DS sind

$$u_i = \Big(\frac{\varepsilon}{\varepsilon + 2h}\Big)^i, \quad i = 0, 1, ..., \infty,$$

$$u_i = \frac{\mu^i - \mu^N}{1 - \mu}, \quad \mu = \frac{\varepsilon}{\varepsilon + 2h} = \frac{1}{1 + \beta} < 1, \ \beta = \frac{2h}{\varepsilon} = 2R_h, \ i = 0, 1, ..., N.$$

Die Fehlerordnung $\mathcal{O}(\frac{h}{\varepsilon})$ der Lösung u_i ist kleiner als in (1.6). Auch hier ist es sinnvoll, grundsätzlich $h < \varepsilon$ zu wählen. Aber man bemerkt, dass keine oszillierenden Lösungsanteile auftreten können.

Variante (1)+(4): DS mit anderer Wahl der Approximation von U' (entgegen den physikalischen Gesetzen)

$$\begin{aligned} -\varepsilon u_{i+1} + (2\varepsilon - 2h)u_i - (\varepsilon - 2h)u_{i-1} &= 0, \quad i = 1, 2, ..., N(\infty) - 1, \\ u_0 = 1, \ u_\infty = 0 \qquad &\text{bzw.} \quad u_0 = 1, \ u_N = 0. \end{aligned} \tag{1.8}$$

Die exakte Lösung des ersten DS ist

$$u_i = \Big(\frac{2\varepsilon - h(1 + \sqrt{1 + 4\varepsilon/h})}{2\varepsilon}\Big)^i, \quad i = 0, 1, ..., \infty.$$

Die Fehlerordnung $\mathcal{O}(\frac{h}{\varepsilon})$ der Lösung u_i ist wie in (1.7). Auch hier wählt man $h < \varepsilon$. Man beachte, dass im DS zumindest eine Vorzeicheneigenschaft erfüllt sein sollte, nämlich $\varepsilon - 2h > 0$, d. h. $h < \frac{\varepsilon}{2}$.

Das bedeutet eine strengere Forderung an die Maschenweite und auch an die diskrete Reynolds-Zahl $R_h = \frac{h}{\varepsilon} \leq \frac{1}{2}$. Damit wird dieser Approximationszugang praktisch nicht angewendet.

Da die RWA nicht in Divergenzform vorliegt, haben alle drei entstehenden LGS keine symmetrische Koeffizientenmatrix A. Aber es bleiben noch solche guten Eigenschaften wie tridiagonal, irreduzibel diagonaldominant, positiv definit, Vorzeichensituation in Haupt- und Nebendiagonalen sowie nichtnegative Elemente der inversen Matrix (Merkmale einer M-Matrix) erhalten.
Den "unendlichen" Fall kann man auf ein großes endliches System von Differenzengleichungen reduzieren, indem man das Ausgangsintervall $[0, b],\ b \gg 1$, betrachtet. Am rechten Rand muss man eine geeignete Randbedingung definieren, wie z. B. die Näherung $U(b) = 0$ oder man nimmt bei Kenntnis von u_i den "rechten" Wert u_N.

(B) Boxmethode

Ein anderer Weg zur Konstruktion des DS geht über die Betrachtung von Energiebilanzen in sogenannten Boxen (Teilintervallen) von $[0, b]$.
Wir beschränken uns hier auf das Intervall $[0, b] = [0, 1]$.
Die Boxmethode (Finite-Volumen-Schema) in der Box $[x_{i-1}, x_i]$ basiert auf der Differenzengleichung

$$-\frac{\varepsilon}{h}\Big(\frac{u_{i+1} - u_{i-1}}{2h} - \frac{u_i - u_{i-2}}{2h}\Big) - 2\frac{u_i - u_{i-1}}{h} \ = \ 0$$

bzw.

$$-u_{i-2} + \Big(1 + \frac{4h}{\varepsilon}\Big)u_{i-1} + \Big(1 - \frac{4h}{\varepsilon}\Big)u_i - u_{i+1} \ = \ 0, \quad i = 2, 3, ..., N-1, \tag{1.9}$$

mit den Randbedingungen $u_0 = 1,\ u_N = 0$ und der Extraformel am rechten Rand

$$-\frac{\varepsilon}{h}\Big(\frac{u_N - u_{N-1}}{h} - \frac{u_{N-1} - u_{N-2}}{h}\Big) - 2\frac{u_N - u_{N-1}}{h} = 0$$

bzw.

$$-u_{N-2} + \Big(2 + \frac{2h}{\varepsilon}\Big)u_{N-1} = 0.$$

Die exakte Lösung des DS ist

$$u_i \ = \ \frac{\mu_1^i - \mu_1^N + \frac{\alpha}{\gamma}(\mu_2^i - \mu_2^N)}{1 - \mu_1^N + \frac{\alpha}{\gamma}(1 - \mu_2^N)},$$

wobei

$$\mu_{1,2} = \frac{1}{\beta \pm \sqrt{1 + \beta^2}},\ \beta = \frac{2h}{\varepsilon},\ 0 < \mu_1 < 1,\ -1 < \mu_2 < 0,$$

$$\alpha = \Big(\frac{1 - \mu_1}{1 - \mu_2}\Big)^3,\ \gamma = \frac{\mu_1}{\mu_2}.$$

Damit tritt in der Lösung eine oszillierende Komponente auf, die jedoch mit wachsendem Index i abnimmt.
Die Koeffizientenmatrix des LGS ist

$$A(N-1,N-1) = \begin{pmatrix} 1+\frac{4h}{\varepsilon} & 1-\frac{4h}{\varepsilon} & -1 & & & 0 \\ -1 & 1+\frac{4h}{\varepsilon} & 1-\frac{4h}{\varepsilon} & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & 1+\frac{4h}{\varepsilon} & 1-\frac{4h}{\varepsilon} & -1 \\ & & & -1 & 1+\frac{4h}{\varepsilon} & 1-\frac{4h}{\varepsilon} \\ 0 & & & & -1 & 2+\frac{2h}{\varepsilon} \end{pmatrix}.$$

Ohne Konvektionsglied $\frac{4h}{\varepsilon}$ erhält man die Matrix

$$\tilde{A} = \begin{pmatrix} 1 & 1 & -1 & & & 0 \\ -1 & 1 & 1 & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & 1 & 1 & -1 \\ & & & -1 & 1 & 1 \\ 0 & & & & -1 & 2 \end{pmatrix}.$$

und das LGS $\tilde{A}u = (1,0,...,0)^T$.

Eigenschaften der Matrix $\tilde{A}$ sind:
– regulär,
– Bandstruktur,
– nicht symmetrisch,
– nicht diagonaldominant,
– ungünstige Vorzeichensituation in Haupt- und Nebendiagonalen,
– die inverse Matrix hat in der ersten Spalte den Vektor $\frac{1}{N}(N\text{–}1, N\text{–}2, ..., 2, 1)^T$,
– EW genügen der Bedingung $\min\limits_{h\to 0} |\lambda(\tilde{A})| = 0$, somit ist $\|\tilde{A}^{-1}\|$ nicht beschränkt.

Aber die Lösung des LGS $u = \tilde{A}^{-1}(1,0,...,0)^T$ erfüllt die Eigenschaft der Monotonie der Lösung, d. h. $1 = u_0 > u_1 > ... > u_{N-1} > u_N = 0$.
Will man in der Matrix A die gute Vorzeichensituation mit einer positiven Hauptdiagonale und nicht positiven Nebendiagonalen haben, so erfordert dies die Ungleichung $1 - \frac{4h}{\varepsilon} \leq 0$, was eine Beschränkung der Maschenweite h von unten nach sich zieht. Zusammen mit $h \leq \varepsilon$ erhält man ein zulässiges Intervall

$$\frac{\varepsilon}{4} \leq h \leq \varepsilon.$$

Die Maschenweite darf also auch nicht zu klein werden.

(C) Finite-Element-Methode (FEM)

In unserem Fall kann man die DGL mit einer Transformation auf die Divergenzform bringen.

$$-\varepsilon U''(x) - 2U'(x) = 0 \quad \Rightarrow \quad -(K(x)U'(x))' = 0 \ \text{ mit } \ K(x) = e^{2x/\varepsilon} \geq K_0 > 0.$$

Wir betrachten das Problem auf $[0,1]$ und wählen als Diskretisierungsverfahren die FEM bei einem äquidistanten Gitter mit der Maschenweite $h = \frac{1}{N}$. Diese basiert auf einer integralen Formulierung des Problems als Variationsaufgabe und speziellen Ansatz- und Testfunktionen.
Wir nehmen auf den Teilintervallen (Elementen) einfache lineare Ansatzfunktionen (Hütchenfunktionen), die im Standardintervall [0,1] die Gestalt

$$\Phi(\xi) = [\Phi_1(\xi), \Phi_2(\xi)]^T = [1-\xi, \xi]^T, \ \xi \in [0,1],$$

haben. Damit wird für jeweils 2 Unbekannte eine lokale Gleichung formuliert. Dabei ist auf jedem Element der Term

$$k_e = \frac{1}{h} \int_0^1 \tilde{K}(\xi) \Phi'(\xi) \Phi'^T \, d\xi \approx \frac{K_e}{h} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

auszuwerten, der als Elementsteifigkeitsmatrix bezeichnet wird.
$\tilde{K}(\xi)$ ist die Transformierte von $K(x)$ in $[x, x+h]$ auf das Standardintervall $[0,1]$, K_e ist ein Zwischenwert von $\tilde{K}(\xi)$ und damit von $K(x)$ aus dem jeweiligen Intervall.
Die Zusammenfassung (Assemblierung) aller lokalen Gleichungen zu einem Gesamtsystem ergibt

$$Au = \frac{1}{h} \begin{pmatrix} \boxed{\begin{matrix} K_1 & -K_1 \\ -K_1 & K_1+K_2 \end{matrix}} & & & & & 0 \\ & \boxed{\begin{matrix} K_1+K_2 & -K_2 \\ -K_2 & K_2+K_3 \end{matrix}} & & & & \\ & & \ddots & \ddots & \ddots & \\ & & & -K_{N-1} & \boxed{\begin{matrix} K_{N-1}+K_N & -K_N \\ -K_N & K_N \end{matrix}} & \\ 0 & & & & & \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = 0,$$

bzw. mit Einbeziehung der Randbedingungen $u_0 = 1, \ u_N = 0$ die Form

$$\frac{1}{h} \begin{pmatrix} h^2/2 & 0 & & & & 0 \\ 0 & K_1+K_2 & -K_2 & & & \\ & -K_2 & K_2+K_3 & -K_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -K_{N-1} & K_{N-1}+K_N & 0 \\ 0 & & & & 0 & 1 \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} h/2 \\ K_1/h \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

Es gilt $A = A^T > 0$ und A^{-1} hat nur nichtnegative Elemente.
Problematisch ist jedoch die schlechte Kondition der Matrix.

1.3 Temperaturverlauf in einer dünnen quadratischen Platte

Gegeben sei die partielle Differentialgleichung für eine Funktion $u(x, y)$ auf dem Einheitsquadrat, die den Temperaturverlauf in einer dünnen Platte beschreibt.

$$-\Delta U(x,y) = -\left(\frac{\partial U}{\partial x^2} + \frac{\partial U}{\partial y^2}\right) = Q(x,y), \quad (x,y) \in \Omega = (0,1)^2. \tag{1.10}$$

Auf dem Rand des Gebietes sei $U(x, y)$ gleich Null.
Das ist eine elliptische RWA bzw. die Poisson-Gleichung.
Der Diskretisierungsparameter bzw. die Maschenweite des quadratischen Gitters sei $h = 1/(n + 1)$. Man diskretisiert die partiellen Ableitungen mittels zentraler Differenzenquotienten 2. Ordnung

$$\Delta U(x_i, y_j) \approx \frac{1}{h^2}(U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4U_{ij})$$

und notiert die Differenzenformel $4u_{ij} - (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) = h^2 q_{ij}$ (Differenzenstern) für alle inneren (zweidimensionalen) Knoten $(x_i, y_j),\ i, j = 1, 2, ..., n$, in linearer Reihenfolge zeilenweise gemäß $(j - 1)n + i$.
Die diskretisierte RWA oder FDM schreibt man als LGS $Au = h^2 q$.
Welche Struktur und Eigenschaften hat die Matrix A? Wie groß ist ihre Bandbreite?
A besitzt die folgende n-dimensionale Blockstruktur.

$$A = \begin{pmatrix} B & -I & & & \\ -I & B & -I & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -I \\ & & & -I & B \end{pmatrix}$$

mit der $(n \times n)$-Matrix

$$B = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 4 \end{pmatrix} = \text{tridiag}(-1, 4, -1)$$

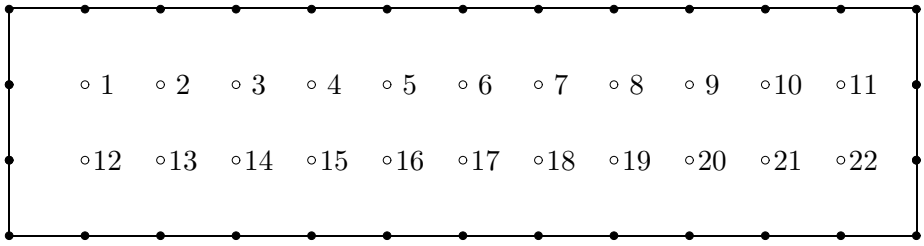
und der $(n \times n)$-Einheitsmatrix I. A ist eine dünn besetzte symmetrische Matrix mit Bandstruktur. Die Bandbreite beträgt $2n + 1$. Die Matrix ist irreduzibel diagonaldominant. EW und EV von A lassen sich aus dem eindimensionalen Fall einfach herleiten.

1.4 Poisson-Gleichung in einem Streifen

Gegeben sei die partielle Differentialgleichung (Poisson-Gleichung) für eine Funktion $U(x, y)$ auf einem Rechteckgebiet.

$$-\Delta U(x,y) = -\left(\frac{\partial U}{\partial x^2} + \frac{\partial U}{\partial y^2}\right) = Q(x,y), \quad (x,y) \in \Omega = (a,b) \times (c,d). \tag{1.11}$$

Auf dem Rand des Gebietes sei die Funktion $U(x, y)$ vorgegeben (Dirichletsche Randbedingungen). Sei $\overline{\Omega} = [a, b] \times [c, d] = [0, 4L] \times [0, L],\ L > 0$.
Zur Lösung verwenden wir die FDM auf einem (13×4)-Gitter (x_i, y_j) mit der Maschenweite $h = L/3$. Wir führen die Nummerierung der $22 = 11 \cdot 2$ unbekannten inneren Punkte im rechteckigen Gitter zeilenweise durch.

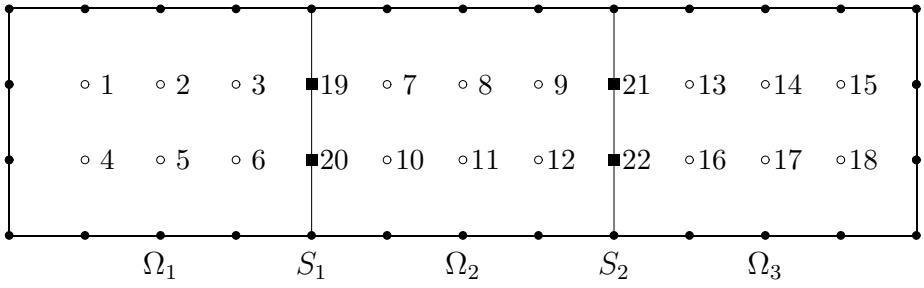


So erhalten wir eine Matrix mit der (2×2)-Blockstruktur

$$A = \begin{pmatrix} B & -I \\ -I & B \end{pmatrix}$$

mit (11×11)-Matrizen $B = \text{tridiag}(-1, 4, -1)$ und I. A hat die Bandbreite 23.

Jetzt nehmen wir eine Gebietszerlegung vor. Folgende Zerlegung von Ω in Teilgebiete Ω_i und "innere Ränder" S_i sei gegeben.



Man diskretisiert die partiellen Ableitungen wiederum mittels zentraler Differenzenquotienten und notiert die Differenzenformel für alle inneren 22 Knoten $(x_i, y_j),\ i = 1, 2, ..., 12,\ j = 1, 2$, in der Reihenfolge 1,2,...,18,19,...,22.
Hierbei beschreiben die inneren Ränder S_i Punkte, welche bei der Diskretisierung die beiden benachbarten breiten Teilgebiete beeinflussen. Sie werden in der Nummerierung als letzte berücksichtigt.

Dies führt auf die folgende Matrixstruktur

$$A = \begin{pmatrix} B & & & C_1 \\ & B & & C_2 \\ & & B & C_3 \\ C_1^T & C_2^T & C_3^T & D \end{pmatrix}$$

mit der (6×6)-Matrix B den (6×4)-Matrizen C_i und der (4×4)-Matrix D.
Natürlich hat auch hier die Matrix A eine sparse Struktur. Dazu kommt die spezielle Blockverteilung der Gestalt eines "Pfeils nach unten"

$$\begin{pmatrix} * & & & * \\ & * & & * \\ & & * & * \\ * & * & * & * \end{pmatrix},$$

die für die Anwendung eines Eliminationsverfahrens und den dabei neu entstehenden Elementen an den bisherigen "Nullstellen" eine günstige Situation darstellt. So kann das sogenannte **Fill-in**, d. h. der Zuwachs der Anzahl der NNE, bei Durchführung des GA gering gehalten werden.

Natürlich können die partiellen Differentialgleichungen wesentlich komplexer und die Problemgebiete viel komplizierter sein.
So hat ein lineares Elastizitätsproblem im $\mathbb{R}^2$ oder $\mathbb{R}^3$ die Gestalt

$$\begin{aligned} -G[\Delta u + \eta \operatorname{grad}(\operatorname{div}(u))] &= f, \quad \Omega \subset \mathbb{R}^d, \ d = 2, 3, \\ u &= u_0 \quad \text{auf } \Gamma_1 \subset \partial\Omega, \\ \sum_{j=1}^{d} \sigma_{ij}\eta_j &= g_i, \quad i = 1, 2, ..., d, \ \text{auf } \Gamma_2 \subset \partial\Omega, \end{aligned}$$

oder ein inkompressibles Elastizitäts- und Strömungsproblem die Gleichungen

$$\begin{aligned} -\operatorname{div}[\eta(v)(\operatorname{grad}(v) + \operatorname{grad}(v^T))] - \rho v \operatorname{grad}(v) - \operatorname{grad}(p) &= f, \\ \operatorname{div}(v) &= 0 \end{aligned}$$

und Randbedingungen.

Auf die Bedeutung und Zusammenhänge bezüglich der auftretenden Variablen und Beziehungen wird hier nicht eingegangen.
Als Diskretisierungsverfahren kommt oft die FEM zur Anwendung, was auf große sparse lineare und nichtlineare Gleichungssysteme führt. Leider gehen aber die anderen günstigen Eigenschaften wie Symmetrie, Definitheit, Bandstruktur, Diagonaldominanz u. a. meist verloren.

1.5 Diskrete Approximation im Mittel

Die folgende Problemstellung wird betrachtet.

• Zwischen zwei Größen x und y wird ein funktionaler Zusammenhang $y = f(x)$ vermutet, der jedoch häufig nur durch Messungen an den Messpunkten $x_0, x_1, ..., x_N$ bekannt ist.

• Eine grafische Darstellung der Messwerte $y_0, y_1, ..., y_N$ über den Messpunkten gibt mitunter Aufschluss über den Typ der zu Grunde liegenden Funktion.
In der Abbildung wird nicht eine lineare Funktion wie im linken Teil, sondern eine quadratische Funktion vermutet, d. h. ein Ansatz der Form $\varphi(x) = a_0 + a_1 x + a_2 x^2$ scheint also sinnvoll.

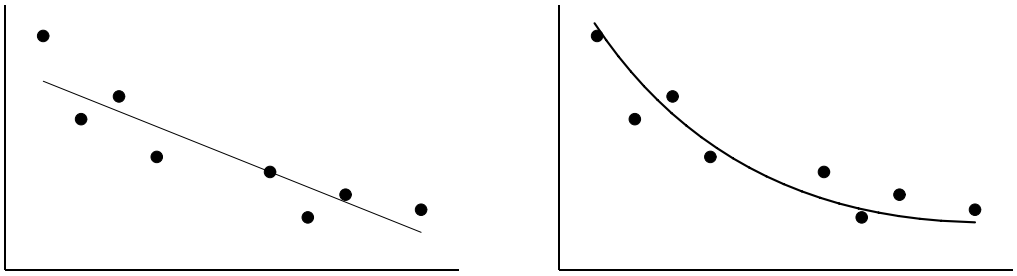


Abb. 1.1 Linearer und quadratischer Ausgleich von Messdaten

• Bei der numerischen Approximation wird eine konkrete Funktion $\varphi(x)$ aus einer vorgegebenen Funktionenklasse so durch die "Punktwolke" der Messdaten (x_j, y_j), $j = 0, 1, ..., N$, hindurchgelegt, dass der **Abstand** zu diesen Punkten insgesamt minimal wird. Der Abstandsbegriff kann sich z. B. auf das Lot, aber auch auf die Entfernung in vertikaler oder horizontaler Richtung beziehen.

• Die Punktwolke kann auch allgemeiner beschrieben sein (mehrdimensionaler Fall).

1.5.1 Die Methode der kleinsten Quadrate

Zur Beschreibung der Methode verwendet man folgende Größen und Bezeichnungen.

- Abgeschlossene, beschränkte Grundmenge $D \subset \mathbb{R}^r$, $r \geq 1$, mit nichtleerem Inneren $\text{int}(D)$ und eine reelle Funktion $f : D \to \mathbb{R}$.

- $N+1$ **Stützstellen** (Messpunkte) x_j und zugehörige **Stützwerte** (Messwerte) y_j sowie als **Referenz**

$$\begin{aligned} R \;=\; \{(x_j, y_j) \mid x_j \;=\; (x_{j1}, x_{j2}, ..., x_{jr}) \in D,\; y_j \in \mathbb{R}, \\ x_j \text{ nicht unbedingt paarweise verschieden},\; j = 0, 1, ..., N\}. \end{aligned} \tag{1.12}$$

  Falls eine Funktion $f(x)$ zu Grunde liegt, definiert man die Stützwerte i. Allg. gemäß $y_j = f(x_j),\; j = 0, 1, ..., N$.

- **Ansatzfunktion** $\varphi(x)$ der Form

$$\varphi(x) \;=\; a_0\varphi_0(x) + a_1\varphi_1(x) + ... + a_n\varphi_n(x), \quad a_i \in \mathbb{R},\; x \in \mathbb{R}^r, \tag{1.13}$$

  mit vorgegebenen $n+1$ **Basisfunktionen** $\varphi_i(x),\; i = 0, 1, ..., n.$

- **Approximationsforderung** (Approximationsbedingung)

$$F \;=\; F(a_0, a_1, ..., a_n) \;=\; \frac{1}{2}\sum_{j=0}^{N}(y_j - \varphi(x_j))^2 \quad \longrightarrow \quad \min, \tag{1.14}$$

  aus der sich auch der Name der Methode herleitet.

- **Approximationsaufgabe (AA)**
  Gesucht sind die Koeffizienten $a_0, a_1, ..., a_n$ in (1.13), so dass die Funktion F minimal wird.

Varianten der Approximation ergeben sich durch verschiedene Ansatzfunktionen.

(1) Die allgemeine lineare Approximation hat die Gestalt

$$\varphi(x) \;=\; a_0\varphi_0(x) + a_1\varphi_1(x) + ... + a_n\varphi_n(x)$$

mit den Basisfunktionen $\varphi_i(x)$.

(2) Die Approximation durch Polynome in $\mathbb{R}^1$ verwendet den Ansatz

$$\varphi(x) \;=\; a_0 + a_1x + a_2x^2 + ... + a_nx^n, \quad a_i \in \mathbb{R},$$

mit den Basisfunktionen $\varphi_i(x) = x^i$. Spezialfälle sind

$$\begin{aligned} \varphi(x) &= a_0 + a_1x && \text{(linearer Ausgleich)}, \\ \varphi(x) &= a_0 + a_1x + a_2x^2 && \text{(quadratischer Ausgleich)}. \end{aligned}$$

(3) Eine Approximation durch trigonometrische Polynome führt z. B. auf

$$\varphi(x) \;=\; a_0 + a_1\cos(x) + a_2\sin(x).$$

(4) Die lineare Approximation in $\mathbb{R}^n$ bedeutet den Ausgleich durch die Hyperebene

$$\varphi(x_1, x_2, ..., x_n) \;=\; a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n,\; x = (x_1, x_2, ..., x_n).$$

Dann wählen wir zwecks Unterscheidung der Variablen für die Stützstellen die andere Bezeichnung $\bar{x}_j$.

Drei Fragestellungen bzw. Probleme sind bei der Lösung der Approximationsaufgabe von Bedeutung.

1. Existiert zu jeder gegebenen Referenz R mit $N+1$ Punkten eine approximierende Funktion $\varphi(x)$ aus der gegebenen Funktionenklasse?
2. Ist $\varphi(x)$ eindeutig bestimmt?
3. Wie kann $\varphi(x)$ effektiv konstruiert werden?
4. Wie berechnet man den **Approximationsfehler** F_{min}?

Herleitung der Methode der kleinsten Quadrate von C.F. Gauß

Gesucht ist ein Koeffizientensatz $a = (a_0, a_1, ..., a_n)$, für den die Summe der Abweichungsquadrate bzw. die **Ausgleichsfunktion**

$$F(a_0, a_1, ..., a_n) \;=\; \frac{1}{2}\sum_{j=0}^{N}\left[y_j - \sum_{i=0}^{n} a_i\varphi_i(x_j)\right]^2 \tag{1.15}$$

minimal wird.

Die notwendigen Bedingungen für ein (lokales) Minimum sind

$$\frac{\partial F}{\partial a_k} \;=\; \sum_{j=0}^{N}\left[\sum_{i=0}^{n} a_i\varphi_i(x_j) - y_j\right]\varphi_k(x_j) \;=\; 0, \quad k = 0, 1, ..., n.$$

Das Auflösen der Klammern und Umstellen liefern

$$\sum_{i=0}^{n}\left[\sum_{j=0}^{N}\varphi_k(x_j)\varphi_i(x_j)\right]a_i \;=\; \sum_{j=0}^{N}\varphi_k(x_j)y_j, \quad k = 0, 1, ..., n.$$

Durch Einführung der **Gaußschen Klammern** als Skalarprodukte in einem reellen Funktionenraum gemäß

$$\begin{aligned}(\varphi_k, \varphi_i) \;&=\; \sum_{j=0}^{N}\varphi_k(x_j)\varphi_i(x_j),\\ (\varphi_k, y) \;&=\; \sum_{j=0}^{N}\varphi_k(x_j)y_j,\end{aligned} \tag{1.16}$$

lassen sich die $n+1$ Bedingungen für die $n+1$ Unbekannten a_i in der Form der **Normalgleichungen** notieren.

Die Kurzform ist

$$\sum_{i=0}^{n}(\varphi_k, \varphi_i)a_i \;=\; (\varphi_k, y), \quad k = 0, 1, ..., n, \tag{1.17}$$

die Langform

$$\begin{aligned}
(\varphi_0,\varphi_0)a_0 + (\varphi_0,\varphi_1)a_1 + ... + (\varphi_0,\varphi_n)a_n &= (\varphi_0, y),\\
(\varphi_1,\varphi_0)a_0 + (\varphi_1,\varphi_1)a_1 + ... + (\varphi_1,\varphi_n)a_n &= (\varphi_1, y),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
(\varphi_n,\varphi_0)a_0 + (\varphi_n,\varphi_1)a_1 + ... + (\varphi_n,\varphi_n)a_n &= (\varphi_n, y).
\end{aligned}$$

Unabhängig von den vorliegenden Daten weist die symmetrische Koeffizientenmatrix G stets eine weitgehend voll besetzte Struktur auf.
Für die Matrix G heißt die Determinante

$$\det(G) \;=\; \det(G(\varphi_0,\varphi_1,...,\varphi_n)) \;=\; \det((\varphi_i,\varphi_j)_{i,j=0}^{n}) \tag{1.18}$$

auch **Gramsche Determinante**.

Definition 1.1 Verallgemeinerte Haarsche Bedingung

Die $n+1$ Basisfunktionen $\varphi_i(x),\ i = 0, 1, ..., n$, erfüllen die verallgemeinerte Haarsche Bedingung, falls für jede Stützstellenfolge in D die **Haarsche Matrix**

$$\Phi \;=\; \begin{pmatrix}
\varphi_0(x_0) & \varphi_1(x_0) & \varphi_2(x_0) & \cdots & \varphi_n(x_0)\\
\varphi_0(x_1) & \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_n(x_1)\\
\varphi_0(x_2) & \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_n(x_2)\\
\cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots\\
\varphi_0(x_N) & \varphi_1(x_N) & \varphi_2(x_N) & \cdots & \varphi_n(x_N)
\end{pmatrix} \tag{1.19}$$

den maximalen Rang $r = \min(n, N) + 1$ besitzt.
Im Fall $n = N$ ist $\det(\Phi)$ *die* **Haarsche Determinante**.

Satz 1.1 Existenz und Eindeutigkeit der Approximationsaufgabe

Falls $N \geq n$ und die verallgemeinerte Haarsche Bedingung erfüllt sind, liefern die Normalgleichungen (1.17) das absolute Minimum der Ausgleichsfunktion (1.15). Weiterhin ist die Approximationsaufgabe für beliebige Stützstellenfolgen in D stets eindeutig lösbar.

Beweis. [28]

Bemerkung 1.1 (1) Falls $N \geq n$ ist, müssen zur Sicherung von $\text{rang}(\Phi) = n + 1$ unter den $N + 1$ Stützstellen x_j mindestens $n + 1$ verschiedene sein.

(2) Der Fall $N < n$ führt i. Allg. auf ein unterbestimmtes System (zu wenig Bedingungen) und wird wegen der Mehrdeutigkeit der Lösung nicht weiter betrachtet.

(3) Sei $N = n$. Dann heißt das System von Basisfunktionen $\{\varphi_i(x)\}$, das mit einer beliebigen Referenz mit paarweise verschiedenen Stützstellen eine stets reguläre Matrix Φ bildet, auch **Haarsches**, Tschebyscheff-, T- oder unisolventes System. Somit ist die Aufgabe wie ein Interpolationsproblem eindeutig lösbar.

(4) Sei $N \geq n$. Es ergeben sich analoge Aussagen zu (3) mit einer Referenz, die mindestens $n + 1$ verschiedene Stützstellen enthalten muss. Die Regularität von Φ wird ersetzt durch $\text{rang}(\Phi) = n + 1$ oder $\det(G) \neq 0$.

Hat man ausschließlich das System der Basisfunktionen $\{\varphi_i(x)\}$ im Auge, dann ist i. Allg. nur garantiert, dass es **mindestens eine** Stützstellenfolge gibt, für die Φ nicht singulär wird. Günstiger sind natürlich solche Systeme von Basisfunktionen, welche die Haarsche Bedingung einer stets regulären Matrix Φ erfüllen, unabhängig von der Wahl der paarweise verschiedenen Stützstellen. In [30] findet der Leser einen Überblick über Haarsche Systeme.
Ein Haarsches Funktionensystem ist $\{1, x, ..., x^n\}$ auf $[a, b] \subset \mathbb{R},\ a < b$.

1.5.2 Ausgleich durch Polynome in $\mathbb{R}^1$

Hierbei wenden wir das Haarsche Funktionensystem $\{1, x, ..., x^n\}$ an und fassen die entsprechenden Schritte kurz zusammen.

Die **skalare Ansatzfunktion** ist

$$\varphi(x) \ = \ a_0 + a_1 x + a_2 x^2 + ... + a_n x^n, \quad a_i \in \mathbb{R}, \tag{1.20}$$

mit den Basisfunktionen $\varphi_i(x) = x^i,\ i = 0, 1, ..., n$.

Spezialfälle sind

$$\begin{aligned}
\varphi(x) &= a_0 + a_1 x && \text{(linearer Ausgleich)},\\
\varphi(x) &= a_0 + a_1 x + a_2 x^2 && \text{(quadratischer Ausgleich)},\\
\varphi(x) &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 && \text{(kubischer Ausgleich)}.
\end{aligned}$$

Die **Normalgleichungen** schreibt man wiederum mit den Gaußschen Klammern

$$\begin{aligned}
(\varphi_k, \varphi_i) &= \sum_{j=0}^{N} x_j^k x_j^i \ = \ \sum_{j=0}^{N} x_j^{k+i},\\
(\varphi_k, y) &= \sum_{j=0}^{N} x_j^k y_j, \quad i, k = 0, 1, ..., n.
\end{aligned} \tag{1.21}$$

Die $n+1$ Gleichungen für die $n+1$ Unbekannten a_i notieren wir in der Matrixform

$$\begin{pmatrix} N+1 & \sum x_j & \sum x_j^2 & \cdots & \sum x_j^n \\ \sum x_j & \sum x_j^2 & \sum x_j^3 & \cdots & \sum x_j^{n+1} \\ \sum x_j^2 & \sum x_j^3 & \sum x_j^4 & \cdots & \sum x_j^{n+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum x_j^n & \sum x_j^{n+1} & \sum x_j^{n+2} & \cdots & \sum x_j^{2n} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum y_j \\ \sum x_j y_j \\ \sum x_j^2 y_j \\ \cdots \\ \sum x_j^n y_j \end{pmatrix}. \quad (1.22)$$

Als Spezialfall des Satzes 1.1 formulieren wir den Satz zur Existenz und Eindeutigkeit für den Polynomausgleich.

Satz 1.2 *Falls $N \geq n$ ist und mindestens $n+1$ Stützstellen x_j verschieden sind, so liefert das Normalgleichungssystem (1.22) das absolute Minimum der Ausgleichsfunktion (1.15).*
Die Approximationsaufgabe ist für beliebige Intervalle $\mathrm{I} = [a, b]$, $a < b$, und beliebige Stützstellenfolgen in I mit mindestens $n+1$ verschiedenen Stützstellen stets eindeutig lösbar.

Beweis. [28]

Lösung des Normalgleichungssystems

Zunächst beachte man, dass die Kondition der voll besetzten Koeffizientenmatrix G durchaus schlecht werden kann und $\det(G) = \det(\Phi^T \Phi)$ nahe Null ist für dicht beieinander liegende Stützstellen.

Im allgemeinen Fall sind zum Aufstellen des LGS $(n+1)(n+2)$ Skalarprodukte auszuwerten. Dazu nutzt man das **Falksche Schema**.

					1	x_N	x_N^2	$\cdots$	x_N^n	y_N
					1	x_{N-1}	x_{N-1}^2	$\cdots$	x_{N-1}^n	y_{N-1}
					$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
					1	x_1	x_1^2	$\cdots$	x_1^n	y_1
					1	x_0	x_0^2	$\cdots$	x_0^n	y_0
1	1	...	1	1	$N+1$	$\sum x_j$	$\sum x_j^2$	$\cdots$	$\sum x_j^n$	$\sum y_j$
x_N	x_{N-1}	...	x_1	x_0	$\sum x_j$	$\sum x_j^2$	$\sum x_j^3$	$\cdots$	$\sum x_j^{n+1}$	$\sum x_j y_j$
x_N^2	x_{N-1}^2	...	x_1^2	x_0^2	$\sum x_j^2$	$\sum x_j^3$	$\sum x_j^4$	$\cdots$	$\sum x_j^{n+2}$	$\sum x_j^2 y_j$
...	...	...	...	...	...	...	...	...	...	...
x_N^n	x_{N-1}^n	...	x_1^n	x_0^n	$\sum x_j^n$	$\sum x_j^{n+1}$	$\sum x_j^{n+2}$	$\cdots$	$\sum x_j^{2n}$	$\sum x_j^n y_j$

Tab. 1.1 Polynomausgleich mit Falkschem Schema zum Aufstellen des Normalgleichungssystems

Wegen der Symmetrie und gleicher Summen kann man den Aufwand reduzieren, so dass nur $2n + (n + 1) = 3n + 1$ Skalarprodukte bzw. Summen zu bestimmen sind.
Folgendes Rechenschema ist günstiger.

j	x	y	xy	x^2	x^2y	x^3	x^3y	x^4	$\cdots$	x^{2n}
0	x_0	y_0	x_0y_0	x_0^2	$x_0^2y_0$	x_0^3	$x_0^3y_0$	x_0^4	$\cdots$	x_0^{2n}
1	x_1	y_1	x_1y_1	x_1^2	$x_1^2y_1$	x_1^3	$x_1^3y_1$	x_1^4	$\cdots$	x_1^{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	x_N	y_N	x_Ny_N	x_N^2	$x_N^2y_N$	x_N^3	$x_N^3y_N$	x_N^4	$\cdots$	x_N^{2n}
$N+1$	$\sum x_j$	$\sum y_j$	$\sum x_jy_j$	$\sum x_j^2$	$\sum x_j^2y_j$	$\sum x_j^3$	$\sum x_j^3y_j$	$\sum x_j^4$	$\cdots$	$\sum x_j^{2n}$

Tab. 1.2 Auswertung der Gaußschen Klammern im Falkschen Schema

Sonderfälle sind der lineare Ausgleich ($n=1$) und der quadratische Ausgleich ($n=2$).
Man erkennt oft an der Determinante die schlechte Kondition des LGS.
Um zumindest die großen Elemente der Koeffizientenmatrix zu vermeiden, empfiehlt sich eine Koordinatentransformation der Stützstellen x_j.
Damit erhalten wir ein **skaliertes System**.

Beispiel 1.1 Wir bestimmen den Ausgleich zur Versuchsreihe (Messdaten, $N = 6$)

x_j	7	12	17	22	27	32	37
y_j	83.7	72.9	63.2	54.7	47.5	41.4	36.3

Vermutet wird ein quadratischer Zusammenhang $\varphi(x) = a_0 + a_1x + a_2x^2$.
Das Normalgleichungssystem ist

$$\begin{aligned} 7a_0 + \quad 154a_1 + \quad 4\,088a_2 &= 399.7, \\ 154a_0 + \quad 4\,088a_1 + \quad 120\,736a_2 &= 7\,688.9, \\ 4\,088a_0 + 120\,736a_1 + 3\,795\,092a_2 &= 186\,054.3. \end{aligned}$$

Die Determinante der Matrix beträgt $\det(G) = 2.572\text{E}8$.
Lösung und Ausgleichsfunktion ist

$$\varphi(x) = \mathbf{100.791\,141}\,2 - \mathbf{2.606\,618}\,857x + \mathbf{0.023\,380\,94}8\,02x^2,$$

wobei die Exaktheit der Rechnung im Computeralgebrasystem (CAS) Maple mit dem Standard-GPF *real* (`Digits:=10`) durch die fett hervorgehobenen Mantissenstellen dargestellt ist.

Die mittlere Abweichung ist $F_{min} = \frac{1}{2}\sum_{j=0}^{N}(y_j - \varphi(x_j))^2 = \frac{1}{120}$.

Wir nehmen nun eine Koordinatentransformation der Stützstellen x_j vor und erhalten ein skaliertes System.
Wir schlagen vor, das Intervall $[a, b]$ mit $a = \min x_j$ und $b = \max x_j$ auf das Intervall $[-2, 2]$ zu transformieren, um damit die Größen x_j in der Nähe von ± 1 zu lokalisieren und die Skalarprodukte $\sum_j x_j^k$ in der Größenordnung von 1 oder -1 zu halten.
Die allgemeine Transformation ist

$$\xi(x) \;=\; \frac{4}{b-a}x - 2\frac{b+a}{b-a},$$

die mit den konkreten Werten $a = 7,\ b = 37$ die Funktionen

$$\xi(x) = \frac{2}{15}(x - 22), \quad x = \frac{15}{2}\xi + 22$$

liefert.
Die tansformierten Stützstellen, die nun symmetrisch zur Null liegen, mit den zugehörigen Stützwerten sind

ξ_j	-2	$-\frac{4}{3}$	$-\frac{2}{3}$	0	$\frac{2}{3}$	$\frac{4}{3}$	2
y_j	83.7	72.9	63.2	54.7	47.5	41.4	36.3

Der neue Ansatz als Polynom zweiten Grades sei $\psi(\xi) = b_0 + b_1\xi + b_2\xi^2$.
Das skalierte Normalgleichungssystem hat damit die Gestalt

$$\begin{aligned}
7b_0 + \quad 0 + \quad \frac{112}{9}b_2 \;&=\; 399.7,\\
0 + \frac{112}{9}b_1 + \quad 0 \;&=\; -147.2 - \frac{1}{15},\\
\frac{112}{9}b_0 + \quad 0 + \frac{3136}{81}b_2 \;&=\; 732.4\ .
\end{aligned}$$

Die Determinante seiner Koeffizientenmatrix beträgt 1445, im Vergleich zu $\det(G)$ ein moderater Wert. Daraus ergeben sich mit einer Rechnung bei gleicher GPA die Ausgleichsfunktionen

$$\begin{aligned}
\psi(\xi) &= 54.761\,904\,71 - 11.833\,928\,58\xi + 1.315\,178\,60\xi^2,\\
\bar\varphi(x) &= \psi(\xi(x)) = b_0 + b_1\xi(x) + b_2\xi(x)^2\\
&= \mathbf{100.791\,143}\,1 - \mathbf{2.606\,619\,0}71x + \mathbf{0.023\,380\,952}\,89x^2,
\end{aligned}$$

bei erneuter Kennzeichnung der gültigen Mantissenstellen. Die Ausgleichsfunktion $\bar\varphi(x)$ ist geringfügig genauer. Zum Vergleich sei noch das exakte Ausgleichspolynom

$$100.791\,\overline{142\,857} - 2.606\,\overline{619\,047}x + \ 0.023\,\overline{380\,952}x^2$$

angegeben, dessen Koeffizienten in den Nachkommastellen eine Periode besitzen.

1.6 Splineinterpolation in $\mathbb{R}^1$

Die Konstruktion von Splines beruht auf einer intervallweisen Interpolation, wo man stückweise Polynome niedrigen Grades zu einer glatten Gesamtfunktion zusammensetzt. Die Ausgangssituation wird durch die folgenden Aspekte beschrieben.

- Grundintervall $\mathrm{I} = [a, b] \subset \mathbb{R}$ mit $-\infty < a < b < \infty$ und (bekannte oder unbekannte) reelle Funktion $f : \mathrm{I} \to \mathbb{R}$.

- Gegeben sind Stützstellen x_k, Stützwerte y_k und Schrittweiten $h_k = x_{k+1} - x_k > 0$ sowie $R_0 = \{(x_i, y_i) \mid a \le x_0 < x_1 < x_2 < ... < x_n \le b\}$ als **Referenz** mit $n+1$ paarweise verschiedenen **Stützstellen** und den $n+1$ zugehörigen **Stützwerten**. Falls eine Funktion $f(x)$ zu Grunde liegt, definiert man die Stützwerte i. Allg. gemäß $y_k = f(x_k),\ k = 0, 1, ..., n$.

- Man konstruiert die Splinefunktion als zusammengesetztes Polynom $s(x) = s(x, R_0)$ vom Grad m ($m \ge 1$) mit den folgenden Eigenschaften.

  (a) $s(x)$ ist ein Polynom vom Grad $\le m$ auf jedem der Teilintervalle, d. h. $s(x) \in \mathcal{S}_m(R_0)$ und

  $$s(x) \;=\; s^{(k)}(x) \;=\; \alpha_{k0} + \alpha_{k1}x + ... + \alpha_{km}x^m, \quad x \in [x_k, x_{k+1}], \tag{1.23}$$

  (b) Bezüglich der Glattheit fordert man $s(x) \in \mathcal{C}^{m-1}(\mathrm{I})$.

- Interpolationsforderung (Interpolationsbedingung)

  (1) $$s(x_k) \;=\; y_k \;=\; f(x_k), \quad k = 0, 1, ..., n. \tag{1.24}$$

  (2) An den inneren Punkten $x_1, x_2, ..., x_{n-1}$ ist $s(x)$ stetig differenzierbar bis zur Ordnung $m-1$, d. h., für $k = 1, 2, ..., n-1$ gelten

  $$\begin{aligned} s^{(k-1)}(x_k) &= s^{(k)}(x_k), \\ s^{(k-1)}(x_k)' &= s^{(k)}(x_k)', \\ &\dots \\ s^{(k-1)}(x_k)^{(m-1)} &= s^{(k)}(x_k)^{(m-1)}. \end{aligned} \tag{1.25}$$

- **Interpolationsaufgabe (IA)**
  Gesucht sind auf den n Teilintervallen insgesamt $n(m+1)$ Koeffizienten α_{kl}, $k = 0, 1, ..., n-1,\ l = 0, 1, ..., m$, so dass die Interpolationsforderung erfüllt ist.

Zunächst stellt man fest, dass man für die $n(m+1) = n+1+mn-1$ Unbekannten nur $n+1+m(n-1) = n+1+mn-m$ Interpolationsbedingungen (1)+(2) zur Verfügung hat. Damit besitzt das Polynom $m-1$ freie Parameter. Zwecks Eindeutigkeit sind diese durch weitere Bedingungen zu binden. Im Einzelnen stellt sich die Situation so dar:

$m = 1$: kein freier Parameter, eindeutige Lösung als Polygonzug, linearer Spline,

$m = 2$: 1 freier Parameter, quadratische Splines,

$m = 3$: 2 freie Parameter, kubische Splines.

1.6.1 Einfache Typen von Splines

(A) Linearer Spline

Für $x \in [x_k, x_{k+1}]$ macht man den Ansatz $s^{(k)}(x) = a_k + b_k(x - x_k),\ k = 0, 1, ..., n-1$.
Die $2n$ Bedingungen sind

$$\begin{aligned} s^{(k)}(x_k) &= f_k,\ k = 0, 1, ..., n-1, \quad s^{(n-1)}(x_n) = f_n, \\ s^{(k)}(x_k) &= s^{(k-1)}(x_k), \quad k = 1, 2, ..., n-1. \end{aligned} \tag{1.26}$$

Das Ergebnis auf dem Teilintervall ist die Newtonsche bzw. Lagrangesche Form

$$\begin{aligned} s^{(k)}(x) &= f_k + \frac{f_{k+1} - f_k}{x_{k+1} - x_k}(x - x_k), \\ &= \frac{x_{k+1} - x}{x_{k+1} - x_k} f_k + \frac{x - x_k}{x_{k+1} - x_k} f_{k+1}, \quad k = 0, 1, ..., n-1. \end{aligned} \tag{1.27}$$

(B) Quadratische Splines

Für $x \in [x_k, x_{k+1}]$ sei $s^{(k)}(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2,\ k = 0, 1, ..., n-1$.
Die $3n$ Bedingungen sind

$$\begin{aligned} s^{(k)}(x_k) &= f_k,\ k = 0, 1, ..., n-1, \quad s^{(n-1)}(x_n) = f_n, \\ s^{(k)}(x_k) &= s^{(k-1)}(x_k), \quad k = 1, 2, ..., n-1, \\ s^{(k)}(x_k)' &= s^{(k-1)}(x_k)',\ k = 1, 2, ..., n-1, \\ s^{(0)}(x_0)' &= m_0 \quad (m_0 \text{ gegeben oder approximiert}). \end{aligned} \tag{1.28}$$

Anstelle der letzten Bedingung $s^{(0)}(x_0)' = m_0$ sind auch andere möglich. Sie werden auch Endbedingungen genannt, falls sie am Ende des untersuchten Bereichs definiert werden. Andere Zusatzbedingungen sind

- $s^{(k)}(\bar{x}) = \bar{y}$, $\bar{x}$ ist eine zusätzliche Stelle,
- $s^{(n-1)}(x_n)' = m_n$,
- $s^{(0)}(x_0)' = s^{(n-1)}(x_n)'$, Periodizität verbunden mit $f_0 = f_n$,
- $s^{(0)}(x_0)' = -s^{(n-1)}(x_n)'$, Antiperiodizität verbunden mit $f_0 = f_n$,
- $K(s) = \int_{x_0}^{x_n} \omega(x)[s''(x)]^2 dx \to \min$, $\omega(x) > 0$ gegebene Gewichtsfunktion,

  damit wird die Gesamtkrümmung der Kurve minimal.

Untersuchungen zur Approximationsgüte linearer und quadratischer Splines findet man in [30].

Zwei Varianten zur Bestimmung der quadratischen Splinefunktion

(1) Man berechnet die Splinekoeffizienten als Lösung eines LGS.

Wir verwenden den Ansatz mit der Normalform des Polynoms 2. Grades

$$\begin{aligned} s^{(k)}(x) &= a_k + b_k x + c_k x^2, \quad x \in [x_k, x_{k+1}],\ k = 0, 1, ..., n-1, \\ s^{(k)}(x)' &= b_k + 2c_k x. \end{aligned}$$

Die Zusammenfassung der Interpolationsbedingungen ergibt die folgende Übersicht.

Bedingungen		Anzahl
$s^{(0)}(x_0)' = m_0$		1
$s^{(k)}(x_k) = f_k,$	$k = 0, 1, ..., n-1$	n
$s^{(k)}(x_{k+1}) = f_{k+1},$	$k = 0, 1, ..., n-1$	n
$s^{(k-1)}(x_k)' = s^{(k)}(x_k)',$	$k = 1, 2..., n-1$	$n-1$
		$\Sigma = 3n$

Somit entsteht ein reguläres LGS $A\alpha = \beta$ mit Blockstruktur, im Fall $n = 4$ ist das

$$\begin{pmatrix} 0 & 1 & 2x_0 & & & & & & & & & 0 \\ 1 & x_0 & x_0^2 & & & & & & & & & \\ 1 & x_1 & x_1^2 & & & & & & & & & \\ 0 & 1 & 2x_1 & 0 & -1 & -2x_1 & & & & & & \\ & & & 1 & x_1 & x_1^2 & & & & & & \\ & & & 1 & x_2 & x_2^2 & & & & & & \\ & & & 0 & 1 & 2x_2 & 0 & -1 & -2x_2 & & & \\ & & & & & & 1 & x_2 & x_2^2 & & & \\ & & & & & & 1 & x_3 & x_3^2 & & & \\ & & & & & & 0 & 1 & 2x_3 & 0 & -1 & -2x_3 \\ & & & & & & & & & 1 & x_3 & x_3^2 \\ 0 & & & & & & & & & 1 & x_4 & x_4^2 \end{pmatrix} \begin{pmatrix} a_0 \\ b_0 \\ c_0 \\ a_1 \\ b_1 \\ c_1 \\ a_2 \\ b_2 \\ c_2 \\ a_3 \\ b_3 \\ c_3 \end{pmatrix} = \begin{pmatrix} m_0 \\ f_0 \\ f_1 \\ 0 \\ f_1 \\ f_2 \\ 0 \\ f_2 \\ f_3 \\ 0 \\ f_3 \\ f_4 \end{pmatrix}.$$

Das Gleichungssystem hat eine diagonale Blockstruktur. Durch Zeilenvertauschungen und -transformationen kann man die Koeffizientenmatrix leicht auf eine obere Dreiecksmatrix bringen und zugleich erkennen, dass ihre Diagonalelemente nicht verschwinden. Somit ist die Matrix regulär.

(2) Die sukzessive Berechnung der Parabelstücke ist eine zweite Möglichkeit.

Man führt die zusätzlichen Unbekannten $s'(x_k) = d_k$ ein. Damit haben die Interpolationsbedingungen für das Teilpolynom $s^{(k)}(x) = a_k + b_k x + c_k x^2$ die Gestalt

$$\begin{aligned} s^{(k)}(x_j) &= f_j, \quad j = k, k+1, \\ s^{(k)}(x_k)' &= f_k' = d_k. \end{aligned}$$

Die Transformation des Intervalls $[x_k, x_{k+1}]$ auf das Standardbezugsintervall $[0,1]$ ergibt

$$\begin{aligned} h_k &= x_{k+1} - x_k, \quad x = x_k + t h_k,\ t \in [0,1], \\ s^{(k)}(x) &= s^{(k)}(\, x_k + t h_k) = q^{(k)}(t), \\ q^{(k)}(t) &= \tilde{a}_k + \tilde{b}_k t + \tilde{c}_k t^2, \quad q^{(k)}(t)' = \tilde{b}_k + 2\tilde{c}_k t, \\ s^{(k)}(x_k) &= q^{(k)}(0) = \tilde{a}_k, \\ s^{(k)}(x_{k+1}) &= q^{(k)}(1) = \tilde{a}_k + \tilde{b}_k + \tilde{c}_k, \end{aligned}$$

$$s^{(k)}(x_k)' = \left.\frac{ds^{(k)}(x)}{dx}\right|_{x=x_k} = \left.\frac{ds^{(k)}(x_k + t h_k)}{dt}\right|_{t=0} \frac{dt}{dx} = \left.\frac{dq^{(k)}(t)}{dt}\right|_{t=0} \frac{1}{h_k} = \frac{q^{(k)}(0)'}{h_k} = \frac{\tilde{b}_k}{h_k}.$$

Die Koeffizienten von $q^{(k)}(t)$ ergeben sich aus den Bedingungen

$$\begin{aligned} \tilde{a}_k &= f_k, \\ \tilde{b}_k &= h_k f_k' = h_k d_k, \\ \tilde{a}_k + \tilde{b}_k + \tilde{c}_k &= f_{k+1} \quad \Rightarrow \quad \tilde{c}_k = f_{k+1} - f_k - h_k d_k, \end{aligned}$$

die wir nun in der Darstellung von $s^{(k)}(x)$ anwenden.

$$\begin{aligned} s^{(k)}(x) &= s^{(k)}(x_k + t h_k) = q^{(k)}(t) \\ &= f_k + h_k d_k t + (f_{k+1} - f_k - h_k d_k) t^2 \\ &= f_k + h_k d_k t + h_k \Bigg(\underbrace{\frac{f_{k+1} - f_k}{h_k}}_{g_k \text{ (Steigung)}} - d_k \Bigg) t^2 \\ &= f_k + h_k d_k t + h_k (g_k - d_k) t^2. \end{aligned}$$

Die Anwendung der Stetigkeitsbedingungen für die 1. Ableitung

$$s^{(k-1)}(x_k)' = s^{(k)}(x_k)', \quad k = 1, 2, ..., n-1,$$

liefert

$$\begin{aligned} s^{(k-1)}(x_{k-1} + 1 \cdot h_{k-1})' &= s^{(k)}(x_k + 0 \cdot h_k)', \\ \frac{q^{(k-1)}(1)'}{h_{k-1}} &= \frac{q^{(k)}(0)'}{h_k}, \\ \frac{\tilde{b}_{k-1} + 2\tilde{c}_{k-1}}{h_{k-1}} &= \frac{\tilde{b}_k}{h_k}. \end{aligned}$$

Das Einsetzen von $\tilde{b}_k,\ \tilde{b}_{k-1},\ \tilde{c}_{k-1}$ führt zu

$$\begin{aligned}\frac{1}{h_{k-1}}\left[h_{k-1}d_{k-1}+2h_{k-1}(g_{k-1}-d_{k-1})\right] &= \frac{h_k d_k}{h_k},\\ d_{k-1}+2(g_{k-1}-d_{k-1}) &= d_k,\\ d_{k-1}+d_k &= 2g_{k-1},\quad k=1,2,...,n-1.\end{aligned}$$

Jetzt brauchen wir die Zusatzbedingung, die gemäß $s^{(0)}(x_0)' = f_0' = d_0$ den Wert d_0 bereitstellt. Damit können mit der aufsteigenden Rekursion

$$d_k \;=\; 2g_{k-1}-d_{k-1},\quad k=1,2,...,n-1,$$

die fehlenden Größen d_k ermittelt werden. Endlich haben wir

$$\begin{aligned}s^{(k)}(x) &= f_k+h_k d_k t+(g_k-d_k)h_k t^2 \ \text{ mit } \ t=\frac{x-x_k}{h_k}\\ &= f_k+d_k(x-x_k)+\frac{g_k-d_k}{h_k}(x-x_k)^2.\end{aligned}$$

Das Ergebnis zeigt große Ähnlichkeit zur Newton-Interpolationsformel bei gleichen Stützstellen und mit der 1. und 2. Steigung, die vergleichbar mit d_k bzw. $(g_k-d_k)/h_k$ sind.

Andere Zusatzbedingungen sind auf ähnliche Weise zu behandeln, wobei bei Periodizität und Antiperiodizität mit $d_0 = \pm d_n$ für die Lösung der Interpolationsaufgabe eine Fallunterscheidung bez. n gerade/ungerade erforderlich ist, aber das eventuell zu lösende LGS auch eine einfache Struktur hat.

Ähnliche schwach besetzte oder tridiagonale Matrixstrukturen erhält man bei kubischen Splines.

(C) Kubische Splines

Eine natürliche kubische Splinefunktion $s(x)$ zur Referenz $R_0,\ y_k = f(x_k)$, ist eine reelle Funktion mit folgenden drei Eigenschaften.

(a) $s(x)$ ist in jedem Teilintervall $[x_k, x_{k+1}],\ k=0,1,...,n-1$, ein Polynom höchstens 3. Grades.
(b) $s(x)$ ist in den Intervallen $(-\infty, x_0)$ und (x_n, ∞) ein Polynom 1. Grades.
Das heißt, dass die Krümmung von $s(x)$ an den Stellen x_0, x_n Null ist.
(c) $s(x), s'(x), s''(x)$ sind stetig in $\mathbb{R}$ und $s(x)$ interpoliert $f(x)$ an den $n+1$ Stützstellen x_k.

Darstellung des Splines $s(x)$

Sei $s(x) = s^{(k)}(x)$ für $x \in [x_k, x_{k+1}]$ mit

$$s^{(k)}(x) = a_k + b_k(x-x_k) + c_k(x-x_k)^2 + d_k(x-x_k)^3,\quad k=0,1,...,n-1.$$

Formulierung der Bedingungen mit $f_k = f(x_k)$

$$\begin{array}{lrcll}
(a) & s^{(k)}(x_k) & = & f_k, & k = 0, 1, ..., n, \\
(b) & s^{(k)}(x_k) & = & s^{(k-1)}(x_k), & k = 1, 2, ..., n, \\
 & s^{(k)}(x_k)' & = & s^{(k-1)}(x_k)', & k = 1, 2, ..., n, \\
 & s^{(k)}(x_k)'' & = & s^{(k-1)}(x_k)'', & k = 1, 2, ..., n, \\
(c) & s^{(0)}(x_0)'' & = & 0 & \text{(2 Zusatzbedingungen)}, \\
 & s^{(n)}(x_n)'' & = & 0, &
\end{array}$$

und mit der zusätzlichen Funktion auf $[x_n, \infty)$

$$s^{(n)}(x) = a_n + b_n(x - x_n) + c_n(x - x_n)^2.$$

Die Funktion $s^{(n)}(x)$ wurde künstlich hinzugefügt, ohne die Aufgabenstellung zu verändern, so dass die Anzahl der unbekannten Koeffizienten gleich der Anzahl der Bedingungen $4n + 3$ beträgt.

Weitere Typen kubischer Splines

- Eingespannter Spline (clamped spline)

  $$s'(x_0) = m_0, \quad s'(x_n) = m_n \quad (m_0, m_n \text{ gegeben}).$$

- Periodischer Spline

  $$s'(x_0) \; = \; s'(x_n), \quad s''(x_0) \; = \; s''(x_n), \text{ wobei } f_0 = f_n \text{ sinnvoll ist.}$$

- Spline mit Not-a-knot-Bedingung
  $s(x)$ ist auf $[x_0, x_1]$ und $[x_1, x_2]$ sowie $[x_{n-2}, x_{n-1}]$ und $[x_{n-1}, x_n]$ identisch. Damit erweisen sich die Knoten x_1 und x_{n-1} als überflüssig ("keine eigentlichen Knoten").

Berechnung der Koeffizienten des natürlichen kubischen Splines

1. Rechte Seiten der Bestimmungsgleichungen
Seien die Schrittweiten $h_k = x_{k+1} - x_k$ und

$$e_k = 3\left(\frac{f_{k+1} - f_k}{h_k} - \frac{f_k - f_{k-1}}{h_{k-1}}\right), \quad k = 1, 2, ..., n-1.$$

2. Die Bestimmungsgleichungen für c_k sind

$$h_{k-1}c_{k-1} + 2(h_{k-1} + h_k)c_k + h_k c_{k+1} = e_k, \quad k = 1, 2, ..., n-1,$$

wobei $c_0 = c_n = 0$.

Dies ist ein LGS mit einer irreduzibel diagonaldominanten Tridiagonalmatrix

$$\begin{pmatrix} 2(h_0+h_1) & h_1 & & & 0 \\ h_1 & 2(h_1+h_2) & h_2 & & \\ \ddots & \ddots & \ddots & & \\ & h_{n-3} & 2(h_{n-3}+h_{n-2}) & h_{n-2} & \\ 0 & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-2} \\ c_{n-1} \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n-2} \\ e_{n-1} \end{pmatrix}.$$

3. Die restlichen Splinekoeffizienten ergeben sich zu

$$\begin{aligned} a_k &= f_k, \quad k = 0, 1, ..., n, \\ b_k &= \frac{1}{h_k}(f_{k+1} - f_k) - \frac{1}{3}(2c_k + c_{k+1})h_k, \quad k = 0, 1, ..., n-1, \\ d_k &= \frac{1}{3h_k}(c_{k+1} - c_k), \quad k = 0, 1, ..., n-1. \end{aligned}$$

4. Algorithmus zur Lösung der Bestimmungsgleichungen für c_k

Das LGS kann mit einer speziellen Variante des GA, dem verkürzten GA (engl.: "chase method", russ.: "metod progonka"), gelöst werden. Dabei sind nur zwei Hilfsvektoren γ und g zu verwenden.

$$\begin{aligned} (1)\quad \gamma_0 &= 1,\ \gamma_1 = 2(h_0 + h_1),\quad g_1 = e_1, \\ (2)\quad \gamma_k &= 2(h_{k-1} + h_k) - \frac{h_{k-1}^2}{\gamma_{k-1}}, \\ g_k &= e_k - \frac{h_{k-1}}{\gamma_{k-1}} g_{k-1}, \quad k = 2, 3, ..., n-1, \\ (3)\quad \gamma_n &= 1,\quad g_n = 0, \\ (4)\quad c_n &= 0, \\ c_k &= \frac{1}{\gamma_k}(g_k - h_k c_{k+1}), \quad k = n-1, n-2, ..., 1, \\ c_0 &= 0. \end{aligned}$$

Herleitung der Beziehungen für a_k, b_k, c_k, d_k

- Einsetzen in die Interpolationsbedingungen (a)−(c)

$$\begin{aligned} (a)\quad a_k &= f_k, && k = 0, 1, ..., n, \\ (b1)\quad a_k &= a_{k-1} + b_{k-1}h_{k-1} + c_{k-1}h_{k-1}^2 + d_{k-1}h_{k-1}^3, && k = 1, 2, ..., n, \\ (b2)\quad b_k &= b_{k-1} + 2c_{k-1}h_{k-1} + 3d_{k-1}h_{k-1}^2, \\ (b3)\quad 2c_k &= 2c_{k-1} + 6d_{k-1}h_{k-1}, \\ (c)\quad c_0 &= c_n = 0. \end{aligned}$$

- Umstellung

$$
\begin{aligned}
(b3)\quad d_k &= \frac{1}{3h_k}(c_{k+1}-c_k), && k=0,1,...,n-1,\\
&\qquad d_{k-1} \text{ in } (b2),\ (b1) \text{ einsetzen},\\
(b2)\quad b_k &= b_{k-1}+(c_k+c_{k-1})h_{k-1}, && k=1,2,...,n,\\
(b1)\quad b_k &= \frac{1}{h_k}(a_{k+1}-a_k)-\frac{1}{3}(2c_k+c_{k+1})h_k, && k=0,1,...,n-1,\\
(c)\quad c_0 &= c_n = 0.
\end{aligned}
$$

- $(b1)$ in $(b2)$ einsetzen

$$
\begin{aligned}
\frac{1}{h_k}(a_{k+1}-a_k)-\frac{1}{3}(2c_k+c_{k+1})h_k &= \frac{1}{h_{k-1}}(a_k-a_{k-1})-\frac{1}{3}(2c_{k-1}+c_k)h_{k-1}+\\
&\quad +(c_k+c_{k-1})h_{k-1},\\
h_{k-1}\Big(c_k+c_{k-1}-\frac{2}{3}c_{k-1}-\frac{1}{3}c_k\Big)+\frac{h_k}{3}(2c_k+c_{k+1}) &= \frac{1}{h_k}(a_{k+1}-a_k)-\frac{1}{h_{k-1}}(a_k-a_{k-1}),\\
h_{k-1}c_{k-1}+2(h_{k-1}+h_k)c_k+h_kc_{k+1} &= \frac{3}{h_k}(f_{k+1}-f_k)-\frac{3}{h_{k-1}}(f_k-f_{k-1})=e_k,\\
&\quad k=1,2,...,n-1,\\
&\quad c_0=c_n=0.
\end{aligned}
$$

Lösung des Systems mit Tridiagonalmatrix $\Rightarrow$ c_k $\Rightarrow$ $d_k,\ b_k$.

Mit der gezeigten Konstruktion des Splines kann man folgenden Satz formulieren.

Satz 1.3 Existenz und Eindeutigkeit der kubischen Splinefunktion
Zur Referenz R_0 mit paarweise verschiedenen Stützstellen $x_k,\ k=0,1,...,n$, existiert stets genau eine natürliche kubische Splinefunktion.

Für andere Zusatzbedingungen, z. B. bei periodischen Splines, erhalten wir die Matrixstruktur

$$
\begin{pmatrix}
* & * & & & * \\
* & * & * & & \\
 & \ddots & \ddots & \ddots & \\
 & & * & * & * \\
* & & & * & *
\end{pmatrix}.
$$

1.7 Diskrete Fourier-Transformation

Die Fourier-Transformation (FT) spielt in der Signalverarbeitung eine wichtige Rolle. Dabei geht es um die Unterdrückung des Rauschens oder die Verstärkung bzw. Abschwächung von Kontrasten.
Ein diskreter Signalvektor $c = (c_0, c_1, ..., c_{n-1})^T$ (Eingangsdaten) wird transformiert auf einen sogenannten Spektralvektor $\hat{c} = (\hat{c}_0, \hat{c}_1, ..., \hat{c}_{n-1})^T$.
Mit dem Blick auf die kontinuierliche Transformation einer Funktion

$$\hat{f}(s) = \int_{-\infty}^{\infty} f(t) e^{\imath st}\, dt, \quad \imath = \sqrt{-1} \text{ imaginäre Einheit},$$

ergeben sich die Komponenten des Spektralvektors gemäß

$$\hat{c}_j \;=\; \sum_{k=0}^{n-1} c_k \omega^{kj}, \quad j = 0, 1, ..., n-1, \quad \omega = e^{\imath 2\pi/n}.$$

In Matrixschreibweise erhält man

$$\hat{c} \;=\; A_n c,$$

wobei $A_n = (\omega^{kj})_{k,j=0}^{n-1}$ die Fourier-Matrix ist.
A_n ist voll besetzt, regulär, symmetrisch und $A_n^{-1} = \frac{1}{n}(\omega^{-kj})_{kj}$.

Ziel der Betrachtungen soll sein, die Matrix A_n bzw. eine spaltenvertauschte Version dieser durch Multiplikationen von sparsen und blockstrukturierten Matrizen zu generieren.
Wir tun dies beispielhaft für $n = 8$.
Dann ist

$$\omega = e^{\imath 2\pi/8},\ \omega^{kj} = \omega^{kj \bmod 8},\ \omega^0 = \omega^8 = 1,\ \omega^4 = -1,\ \omega^2 = -\imath$$

und

$$A_8 \;=\; \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 & w^4 & \omega^5 & \omega^6 & \omega^7 \\ 1 & \omega^2 & \omega^4 & \omega^6 & 1 & \omega^2 & \omega^4 & \omega^6 \\ 1 & \omega^3 & \omega^6 & \omega & \omega^4 & \omega^7 & \omega^2 & \omega^5 \\ 1 & \omega^4 & 1 & \omega^4 & 1 & \omega^4 & 1 & \omega^4 \\ 1 & \omega^5 & \omega^2 & \omega^7 & \omega^4 & \omega & \omega^6 & \omega^3 \\ 1 & \omega^6 & \omega^4 & \omega^2 & 1 & \omega^6 & \omega^4 & \omega^2 \\ 1 & \omega^7 & \omega^6 & \omega^5 & \omega^4 & \omega^3 & \omega^2 & \omega \end{pmatrix}.$$

Nach Spaltenvertauschung mit dem Permutationsvektor $(1,5,2,6,3,7,4,8)$ erhält man die Matrix

$$\widetilde{A}_8 = \left(\begin{array}{cccc|cccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega^2 & \omega^4 & \omega^6 & \omega & \omega^3 & \omega^5 & \omega^7 \\ 1 & \omega^4 & 1 & \omega^4 & \omega^2 & \omega^6 & \omega^2 & \omega^6 \\ 1 & \omega^6 & \omega^4 & \omega^2 & \omega^3 & \omega & \omega^7 & \omega^5 \\ \hline 1 & 1 & 1 & 1 & \omega^4 & \omega^4 & \omega^4 & \omega^4 \\ 1 & \omega^2 & \omega^4 & \omega^6 & \omega^5 & \omega^7 & \omega & \omega^3 \\ 1 & \omega^4 & 1 & \omega^4 & \omega^6 & \omega^2 & \omega^6 & \omega^2 \\ 1 & \omega^6 & \omega^4 & \omega^2 & \omega^7 & \omega^5 & \omega^3 & \omega \end{array}\right).$$

Sei I_k die Einheitsmatrix der Dimension k. Dann notieren wir die Matrix $\widetilde{A}_8$ stufenweise in der Form

$$\widetilde{A}_8 = \begin{pmatrix} I_4 & I_4 \\ I_4 & -I_4 \end{pmatrix} \begin{pmatrix} I_4 & 0 \\ 0 & D_4 \end{pmatrix} \begin{pmatrix} A_4 & 0 \\ 0 & A_4 \end{pmatrix} = \begin{pmatrix} A_4 & D_4 A_4 \\ A_4 & \omega^4 D_4 A_4 \end{pmatrix},$$

wobei

$$D_4 = \operatorname{diag}(1,\omega,\omega^2,\omega^3), \quad A_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega^2 & \omega^4 & \omega^6 \\ 1 & \omega^4 & 1 & \omega^4 \\ 1 & \omega^6 & \omega^4 & \omega^2 \end{pmatrix},$$

$$\begin{aligned} \widetilde{A}_8 = {} & \begin{pmatrix} I_4 & I_4 \\ I_4 & -I_4 \end{pmatrix} \begin{pmatrix} I_4 & 0 \\ 0 & D_4 \end{pmatrix} \cdot \\ & \cdot \begin{pmatrix} I_2 & I_2 & & \\ I_2 & -I_2 & & \\ & & I_2 & I_2 \\ & & I_2 & -I_2 \end{pmatrix} \begin{pmatrix} I_2 & & & \\ & D_2 & & \\ & & I_2 & \\ & & & D_2 \end{pmatrix} \begin{pmatrix} A_2 & & & \\ & A_2 & & \\ & & A_2 & \\ & & & A_2 \end{pmatrix}, \end{aligned}$$

wobei

$$D_2 = \operatorname{diag}(1,\omega), \quad A_2 = \begin{pmatrix} 1 & 1 \\ 1 & \omega^2 \end{pmatrix}.$$

Kapitel 2

Grundlagen der linearen Algebra

Dieses Kapitel dient der Einführung von zahlreichen grundlegenden Begriffen und Aussagen der linearen Algebra. Ein Teil davon ist Gegenstand der mathematischen Grundlagenausbildung und wird deshalb nur übersichtsartig zusammengestellt.

2.1 Vektoren und Matrizen

Wir geben eine Zusammenfassung einiger wichtiger Definitionen und Sätze über Vektoren und Matrizen aus dem Raum $\mathbb{C}^n$ bzw. $\mathbb{C}^{n,n}$ - speziell auch aus den reellen Vektorräumen - an, ansonsten möchten wir den Leser z. B. auf die Literatur [9], [19], [46] und [47] verweisen.
Wir führen zunächst das Skalarprodukt im Komplexen und Reellen ein, das in vielen Zusammenhängen verwendet wird.

Definition 2.1 Skalarprodukt

Das Skalarprodukt oder innere Produkt von Vektoren aus $\mathbb{C}^n$ ist definiert durch

$$(x,y) \;=\; x^H y \;=\; \bar{x}^T y \;=\; \sum_{i=1}^{n} \bar{x}_i y_i. \tag{2.1}$$

Als erste wichtige Eigenschaften folgen sofort die Beziehungen $(x,y) = \overline{(y,x)}$ sowie für reelle Vektoren $(x,y) = x^T y = y^T x = (y,x)$.

Neben dem Skalarprodukt $x^H y$ zweier Vektoren x, y gibt es noch ihr dyadisches Produkt.

Definition 2.2 Dyadisches Produkt oder Dyade in $\mathbb{R}^n$

Für die Vektoren $x, y \in \mathbb{R}^n,\ x, y \neq 0$, ist dies die Matrix

$$M \;=\; xy^T \;=\; (y_1 x, y_2 x, ..., y_n x) \;=\; \begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \cdots & \cdots & \cdots & \cdots \\ x_ny_1 & x_ny_2 & \cdots & x_ny_n \end{pmatrix}. \tag{2.2}$$

Zwei beliebige Spalten wie auch Zeilen der Matrix M sind linear abhängig.
Deshalb ist sie singulär. Mehr noch, sie hat höchstens den Rang Eins.

Produkte aus n-dimensionalen Matrizen und Vektoren notieren wir im Reellen als

$$\begin{aligned} y &= Ax, \quad y_i = \sum_{j=1}^{n} a_{ij}x_i, \ i = 1, 2, ..., n, \quad \text{d. h. } A : \mathbb{R}^n \to \mathbb{R}^n, \\ y^T &= x^T A, \quad y_j = \sum_{i=1}^{n} x_i a_{ij}, \ j = 1, 2, ..., n, \end{aligned}$$

Produkte und Potenzen von Matrizen gemäß AB, ABC, $A^0 = I$, wobei I die Einheitsmatrix ist, A^k, $A^k B^l$ usw.

Definition 2.3 Bilinearform und quadratische Form in $\mathbb{R}^n$

Bilinearformen mit n-dimensionalen Matrizen und Vektoren sind

$$\begin{aligned} (x, Ay) &= x^T Ay = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i y_j, \\ (Ax, Ay) &= (Ax)^T Ay = x^T A^T Ay = \sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ki}a_{kj}x_i y_j. \end{aligned} \tag{2.3}$$

Ist $x = y$, dann spricht man von quadratischen Formen.

Natürlich ist auch die analoge Definition im Komplexen möglich.

Man achtet dabei also konsequent auf die Notation der Vektoren als Spaltenvektoren sowie die Verwendung von einspaltigen oder einzeiligen Rechteckmatrizen.

Das CAS **Maple** will sich in der Benutzung von Vektoren als Spalten- oder Zeilenvektoren nicht festlegen, sondern sich nach Bedarf und Stellung eines Vektors in der Formel entscheiden. Das hat zur Folge, dass eigentlich Verwirrungen vorprogrammiert sind. **Matlab** ist diesbezüglich konsequenter.
Wir fassen wichtige Erkenntnisse von unterschiedlichen Implementationen in Maple zusammen.

1. So wird das Kommando des Vektortranspostion `transpose(x)` nicht explizit ausgeführt.

2. Im Matrix-Vektor-Produkt `A&*x` ist x ein Spaltenvektor, während x in `x&*A` als Zeilenvektor genommen wird. Stehen jedoch nur Vektoren in der Formel, wie beim Skalarprodukt `transpose(x)&*x` oder der Dyade `x&*transpose(x)`, dann wird x als Spaltenvektor interpretiert. Also ist die Deutung des links stehenden Vektors x in den Produkten `x&*A` und `x&*transpose(x)` verschieden.

3. Eine Mischung der Situationen von Skalar- und Matrix-Vektor-Produkt ergibt sich bei der Betrachtung der quadratischen Form $x^T Ax$.
   Wegen `x&*A` und `A&*x` ist einerseits die Darstellung `x&*A&*x` möglich.
   Aber genauso kann man wegen `A&*x` und `transpose(x)&*x` den Ausdruck `transpose(x)&*A&*x` notieren.

4. Sobald in größeren Formeln Ausdrücke mit Feldern stehen, z. B. Differenzen von Vektoren oder Matrizen, ist zumeist die Einbeziehung der Berechnungsfunktion `evalm` zu empfehlen, manchmal ist es zur Vermeidung von syntaktischen Fehlern sogar notwendig.

5. Bei Differenzen von Matrizen haben wir die Situation, dass `A-A` nicht die Nullmatrix ist, sondern einfach der Wert 0, so dass dann das Ergebnis auch keine Dimensionsabfragen mit `rowdim( )` bzw. `coldim( )` erlaubt.
   Ist A jedoch inhaltlich mit der Einheitsmatrix I identisch, so ist die Differenz `A-I` eine Nullmatrix mit gegebener Dimension.

6. Bei Potenzen von Matrizen haben wir die Situation, dass `A^0` nicht die Einheitsmatrix ist, sondern einfach der Wert 1, so dass dann das Ergebnis auch keine Dimensionsabfragen mit `rowdim( )` bzw. `coldim( )` erlaubt.
   Bildet man jedoch `A^0-I` mit der Einheitsmatrix I, so erinnert sich Maple daran, dass A^0 eigentlich mehr als die Eins darstellt und berechnet richtig als Differenz die Nullmatrix mit gegebener Dimension.

Zur Kontrolle von Ergebnissen werden in manchen Rechnungen auch Feldkomponenten ausgewertet.

Rechnungen in Maple

Definition von Matrizen und Vektoren

```
> m:=5:
  n:=2:

  A:=matrix(m,m,(i,j)->i+j-1);
  A51:=matrix(m,1,[[ 1],[ 2],[ 3],[ 4],[ 5]]);
  A52:=matrix(m,n,[[ 1, 3],
                   [ 2, 2],
                   [ 3, 1],
                   [ 4, 0],
                   [ 5,-1]]);

  x:=vector(m,[1,1,1,1,1]);  # je nach Bedarf Spalten- oder Zeilenvektor
                             # Vektor an Kommas erkennbar

  b:=vector(m,[1,2,3,4,5]);
  c:=vector(m,[1,0,0,0,0]);
```

$$A := \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 6 & 7 \\ 4 & 5 & 6 & 7 & 8 \\ 5 & 6 & 7 & 8 & 9 \end{bmatrix}$$

$$A51 := \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

$$A52 := \begin{bmatrix} 1 & 3 \\ 2 & 2 \\ 3 & 1 \\ 4 & 0 \\ 5 & -1 \end{bmatrix}$$

$$x := [1, 1, 1, 1, 1]$$

$$b := [1, 2, 3, 4, 5]$$

$$c := [1, 0, 0, 0, 0]$$

Transposition

```
> transpose(A51);
  transpose(A52);

  transpose(x);       # keine explizite Angabe des Vektors,
                      # da nicht klar ob Zeilen- oder Spaltenvektor
  evalm(transpose(x));
```

$$[1 \quad 2 \quad 3 \quad 4 \quad 5]$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 0 & -1 \end{bmatrix}$$

$$\text{transpose}(x)$$

$$\text{transpose}(x)$$

Produkte

Matrix∗Matrix, Matrix∗Vektor, Vektor∗Matrix, ...
Kontrolle von Feldkomponenten

```
> evalm(transpose(A52)&*A52);    # A52(1..5,1..2)
  evalm(A52&*transpose(A52));
  d:=evalm(transpose(A52)&*b);   # b(1..5) Spaltenvektor
  d[1], d[2];

  f:=evalm(b&*A52);              # b(1..5) Zeilenvektor
  f[1], f[2];

  B:=matrix(2,1,[[2],
                 [3]]);
  B[1,1];
  B[2,1];
  B[1,2];    # Fehler
```

```
B:=evalm(transpose(A52)&*b); # A52(1..5,1..2), b(1..5) Spaltenvektor
B[1], B[2];
B[1,1];    # Fehler
B[2,1];    # Fehler
B[1,2];    # Fehler

C:=evalm(b&*A52);           # b(1..5) Zeilenvektor, A52(1..5,1..2)
C[1], C[2];
C[1,1];    # Fehler
```

$$\begin{bmatrix} 55 & 5 \\ 5 & 15 \end{bmatrix}$$

$$\begin{bmatrix} 10 & 8 & 6 & 4 & 2 \\ 8 & 8 & 8 & 8 & 8 \\ 6 & 8 & 10 & 12 & 14 \\ 4 & 8 & 12 & 16 & 20 \\ 2 & 8 & 14 & 20 & 26 \end{bmatrix}$$

$$d := [55, 5]$$

$$55,\ 5$$

$$f := [55, 5]$$

$$55,\ 5$$

$$B := \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$2$$

$$3$$

```
Error, 2nd index, 2, larger than upper array bound 1
```

$$B := [55, 5]$$

$$55,\ 5$$

```
Error, array defined with 1 indices, used with 2 indices
Error, array defined with 1 indices, used with 2 indices
Error, array defined with 1 indices, used with 2 indices
```

$$C := [55, 5]$$

$$55,\ 5$$

```
Error, array defined with 1 indices, used with 2 indices
```

Produkte mit Rechteckmatrizen

```
> R:=evalm(transpose(A52)&*A51);   # A52(1..5,1..2), A51(1..5,1..1)
  R[1,1], R[2,1];
  R[1];

  S:=evalm(transpose(A51)&*A52);
  S[1,1], S[1,2];
  S[1];
```

$$R := \begin{bmatrix} 55 \\ 5 \end{bmatrix}$$

$$55,\ 5$$

```
Error, array defined with 2 indices, used with 1 indices
```

$$S := [55\ 5]$$

$$55,\ 5$$

```
Error, array defined with 2 indices, used with 1 indices
```

Quadratische Form $(Ax)^T Ax$

```
> qfd:=evalm(transpose(evalm(A&*b))&*A&*b);
  evalm(transpose(A&*b)&*A&*b);          # Fehler
  evalm(evalm(transpose(A&*b))&*A&*b);   # Fehler
```

$$qfd := 38375$$

```
Error, (in evalm/amperstar) &* is reserved for matrix multiplication
Error, (in evalm/amperstar) &* is reserved for matrix multiplication
```

Quadratische Form $x^T Ax$

```
> qf:=evalm(b&*A&*b);                # A(1..5,1..5)
  qf:=evalm(transpose(b)&*A&*b);
  evalm(transpose(b)&*A&*(b-c));

  evalm(transpose(b-c)&*A&*b);    # Fehler
  evalm(transpose(b-c)&*A&*(b-c));# Fehler, Differenz vorher berechnen

  bmc:=evalm(b-c):
  evalm(transpose(evalm(b-c))&*A&*evalm(b-c));
  evalm(transpose(bmc)&*A&*bmc);
```

$$qf := 1425$$

$$qf := 1425$$

$$1370$$

```
Error, (in linalg[multiply]) expecting a matrix or a vector
Error, (in linalg[multiply]) expecting a matrix or a vector
```

$$1316$$

$$1316$$

Skalarprodukt $x^T x$

```
> evalm(b);
  evalm(c);
  skal:=evalm(transpose(b)&*b);
  evalm(transpose(b-c)&*(b-c));   # Fehler, Differenz vorher berechnen

  bmc:=evalm(b-c);
  evalm(transpose(evalm(b-c))&*evalm(b-c)),
  evalm(transpose(bmc)&*bmc);
```

$$[1,\ 2,\ 3,\ 4,\ 5]$$
$$[1,\ 0,\ 0,\ 0,\ 0]$$
$$skal := 55$$

```
Error, (in linalg[multiply]) expecting a matrix or a vector
```

$$bmc := [0, 2, 3, 4, 5]$$
$$54,\ 54$$

Skalarprodukt (Normquadrat) $z^T z$ und dyadisches Produkt $z z^T$
Vektor z je nach Bedarf Spalten- oder Zeilenvektor

```
> z:=vector(m,[1$m]);
  transpose(z);
  z1:=matrix(5,1,z);
  transpose(z1);

  # Skalarprodukt
  norm2q:=evalm(transpose(z)&*z); # Normquadrat, z wie Spaltenvektor
  eval(transpose(z)&*z);          # nicht verwertbar
  evalf(transpose(z)&*z);         # nicht verwertbar

  # dyadisches Produkt
  dyad:=evalm(z1&*transpose(z1)); # einsichtige Darstellung mit Matrizen
  dyad:=evalm(z&*transpose(z)):   # auch moeglich, z wie Spaltenvektor
  dyad[1,1], dyad[5,5];

  erg1:=evalm(z&*z1);             # (1)-Vektor, z wie Zeilenvektor
  erg1[1];
  evalm(z&*transpose(z1));        # Fehler, Zeilenvektor*Zeilenvektor
  erg2:=evalm(transpose(z)&*z1);  # ? nicht verwertbar
  erg2[1];
```

$$z := [1, 1, 1, 1, 1]$$
$$\mathrm{transpose}(z)$$
$$z1 := \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$
$$[1\ \ 1\ \ 1\ \ 1\ \ 1]$$
$$norm2q := 5$$
$$\mathrm{transpose}(z)\&^* z$$
$$\mathrm{transpose}(z)\&^* z$$
$$dyad := \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$1,\ 1$$
$$erg1 := [5]$$
$$5$$

```
Error, (in linalg[multiply]) non matching dimensions
                             for vector/matrix product
```

$$erg2 := \mathrm{transpose}([5])$$
$$\mathrm{transpose}([5])_1$$

Weitere Produkte

```
> evalm(z);                        # z(1..5)
  evalm(z1);                       # z1(1..5,1..1)

  ma11:=evalm(transpose(z1)&*z);   # (1)-Vektor, z wie Spaltenvektor
  ma11[1];
  ma12:=evalm(transpose(z1)&*z1);  # (1 x 1)-Matrix
  ma12[1,1];
  ma13:=evalm(z&*z1);              # (1)-Vektor, z wie Zeilenvektor
  ma13[1];

  # Problemsituationen
  d1:=evalm(z1&*z);            # Fehler, Spaltenanz.(z1)<>Zeilenanz.(z)
  d2:=evalm(z1&*transpose(z)); # Fehler, muesste Dyade sein
  d3:=evalm(transpose(z1)&*transpose(z));   # Fehler
  d4:=evalm(transpose(z)&*z1);     # nicht verwertbar
```

$$[1,\ 1,\ 1,\ 1,\ 1]$$
$$:= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$
$$ma11 := [5]$$
$$5$$
$$ma12 := [\ 5]$$
$$5$$
$$ma13 := [5]$$
$$5$$

```
Error, (in linalg[multiply]) non matching dimensions for
                             vector/matrix product
Error, (in linalg[multiply]) expecting a matrix or a vector
Error, (in linalg[multiply]) expecting a matrix or a vector
```

$$d4 := \mathrm{transpose}([5])$$

Matrixdifferenzen und Dimension der Matrix

```
> A1:=diag(1,1);
  rowdim(A);
  A2:=matrix(2,2,[[1,2],[3,4]]);
  rowdim(A1);
  evalm(A2-A1);
  rowdim(A2-A1);
```

$$A1 := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$2$$

$$A2 := \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$2$$

$$\begin{bmatrix} 0 & 2 \\ 3 & 3 \end{bmatrix}$$

$$2$$

```
> n:=2;
  II:=array(identity,1..n,1..n);
  evalm(II);
  rowdim(II);
  I0:=array(sparse,1..n,1..n);
  evalm(I0);
  rowdim(I0);
```

$$n := 2$$

$$II := \mathrm{array}(identity, 1 \,..\, 2, 1 \,..\, 2, [\ ])$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$2$$

$$I0 := \mathrm{array}(sparse, 1 \,..\, 2, 1 \,..\, 2, [\ ])$$

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$2$$

```
> evalm(II-II);      # 0-Wert-Matrix
  rowdim(II-II);     # Fehler
  evalm(A1-A1);      # 0-Wert-Matrix
  rowdim(A1-A1);     # Fehler
  evalm(A1-A1-I0);   # 0-Matrix
  rowdim(A1-A1-I0);
  evalm(A1-II);      # 0-Matrix
  rowdim(A1-II);
```

$$0$$

```
Error, (in rowdim) first argument is zero, need zero matrix
```

$$0$$

```
Error, (in rowdim) first argument is zero, need zero matrix
```

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$2$$

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$2$$

Potenzen mit Vektoren

```
> evalm(x), evalm(b), evalm(c);
  evalm(x^0);

  evalm(x^0+x^1);
  evalm(x^0-b^0);
  evalm(x^0-x);    # 1-[1,1,1,1]=[0,0,0,0]
  evalm(x^0-c);
```

$$[1, 1, 1, 1, 1], [1, 2, 3, 4, 5], [1, 0, 0, 0, 0]$$

$$1$$

$$[2, 2, 2, 2, 2]$$

$$0$$

$$[0, 0, 0, 0, 0]$$

$$[0, 1, 1, 1, 1]$$

Potenzen mit Matrizen

```
> evalm(A2), evalm(II);
  evalm(A2^0);
  rowdim(evalm(A2^0));    # Fehler

  evalm(A2^0+A2^1);
  evalm(A2^0-A1^0);
  evalm(A2^0-A2);         # A2^0-A2=1-A2=I-A2
  evalm(A2^0-II);         # A2^0-II=1-II=0-Matrix
```

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$1$$

```
Error, (in rowdim) expecting a matrix
```

$$\begin{bmatrix} 2 & 2 \\ 3 & 5 \end{bmatrix}$$

$$0$$

$$\begin{bmatrix} 0 & -2 \\ -3 & -3 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Definition 2.4 Orthogonalität und weitere Matrixeigenschaften

(1) Zwei nichtverschwindende Vektoren $x, y \in \mathbb{R}^n$ sind orthogonal, wenn ihr Skalarprodukt $(x, y) = x^T y$ Null ist.

(2) Eine Matrix $A \in \mathbb{R}^{n,n}$ ist orthogonal, falls gilt $AA^T = A^T A = I$, I Einheitsmatrix. Damit bilden die Matrixzeilen bzw. -spalten orthogonale Vektoren.
Manchmal spricht man hier auch von Orthonormalität wegen der zusätzlichen Skalierung (Normalisierung) der Vektoren auf die Länge Eins (euklidische Norm).

(3) Eine Matrix $A \in \mathbb{C}^{n,n}$ ist unitär, falls $AA^H = A^H A = I$ gilt.

(4) Eine Matrix $A \in \mathbb{C}^{n,n}$ ist hermitesch, falls die Matrix A identisch ist mit ihrer konjugiert komplexen und transponierten, also $A = A^H = \bar{A}^T$.
Im Reellen erhalten wir die Symmetrie $A = A^T$, ihre Inverse ist auch symmetrisch.

(5) Eine Matrix $A \in \mathbb{C}^{n,n}$ ist normal, falls $AA^H = A^H A$ gilt.
Eine hermitesche Matrix ist auch normal.

(6) Die adjungierte Matrix A^* von $A \in \mathbb{C}^{n,n}$ ist diejenige, für die $(Ax, y) = (x, A^* y)$ gilt. Die Matrix A^* wird somit unter Verwendung des Skalarprodukts indirekt definiert.

Für Matrizen gilt $A^* = A^H$. Damit ist eine hermitesche Matrix selbstadjungiert.

Definition 2.5 Definitheit einer Matrix

Die Matrix $A \in \mathbb{C}^{n,n}$ ist positiv definit (kurz $A > 0$), falls $x^H Ax = (x, Ax) > 0$ für $x \in \mathbb{C}^n$, $x \neq 0$. Positiv semidefinit, negativ definit, negativ semidefinit bedeuten entsprechend $A \geq 0\ \forall x$, $A < 0\ \forall x \neq 0$ und $A \leq 0\ \forall x$.
Im Fall einer echt komplexen Matrix $A = B + \imath C$ macht die Schreibweise $A > 0$ nur Sinn für $A = A^H$ wegen Satz 2.4. Für eine reelle Matrix bedeutet die positive Definitheit $x^T Ax = (x, Ax) > 0$ mit $x \in \mathbb{R}^n$, $x \neq 0$. Im Fall einer reellen, symmetrischen und positiv definiten Matrix schreibt man $A = A^T > 0$ oder auch das Kürzel **spd**.
Die dabei auftretenden Skalarprodukte (x, Ax) und $(x, Ay) = (A^H x, y)$ sind quadratische bzw. Bilinearformen von A.

Beispiel 2.1 Matrizen $A = (a_{ij})$ und ihre Definitheit

(1) $A = \text{diag}(d_1, d_2, ..., d_n)$ ist positiv gdw. alle Diagonalelemente $d_i > 0$ sind.

(2) $A = A^T$ streng diagonaldominant (Definition 2.8) und $a_{ii} > 0 \ \Rightarrow \ A$ spd

$$A(3,3) = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

(3) $A = A^T$ irreduzibel diagonaldominant (Definition 2.12) und $a_{ii} > 0 \ \Rightarrow \ A$ spd

$$A(3,3) = \text{tridiag}(-1, 2, -1)$$

(4) $A = A^T$ schwach diagonaldominant und $a_{ii} > 0 \;\Rightarrow\; A \geq 0$

$$A(3,3) = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

$x^T Ax = 0$ für $x = (1,1,1)^T \neq 0$

(5) A Dreiecksmatrix, streng diagonaldominant und $a_{ii} > 0 \;\Rightarrow\; A > 0$

$$A(2,2) = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$$

$x^T Ax = (x_1 + \frac{1}{2}x_2)^2 + x_1^2 + \frac{11}{4}x_2^2 > 0$ für alle $x = (x_1, x_2)^T \neq 0$

(5) A obere Dreiecksmatrix und $a_{ii} > 0$ (damit hat A die Eigenwerte $\lambda_i = a_{ii}$)

$$A(2,2) = \begin{pmatrix} 1 & \varepsilon \\ 0 & 1 \end{pmatrix}$$

$x^T Ax = (x_1 + \frac{\varepsilon}{2}x_2)^2 + x_2^2(1 - \frac{\varepsilon^2}{4}) > 0$ für alle $x = (x_1, x_2)^T \neq 0$ bei $|\varepsilon| < 2$

(6) A mit reellen EW $\lambda_i > 0 \;\not\Rightarrow\; A > 0$

$$A(2,2) = \begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix}$$

EW $\lambda_{1,2} = \frac{1}{2000}(1439 \pm \sqrt{2070717}) = 1.438\,999\,305\,072..., \; 0.000\,000\,694\,927... > 0,$
$x^T Ax = 0.780x_1^2 + 1.476x_1x_2 + 0.659x_2^2 = -0.01461 < 0$ für $x = (1, -0.9)^T \neq 0$

Aus linear unabhängigen Vektoren $x_i \in \mathbb{R}^n,\; i = 1, 2, ...,$ kann man $k \leq n$ Vektoren auswählen, die dann eine Basis $\mathrm{X} = \{x_1, x_2, ..., x_k\}$ des linearen Unterraums

$$\mathcal{X} = [\mathrm{X}] = \operatorname{span} \mathrm{X} = \operatorname{span}\{x_1, x_2, ..., x_k\} \subset \mathbb{R}^n, \tag{2.4}$$

bilden. Wir notieren hierbei verschiedene gebräuchliche Bezeichnungen des Unterraums. Seine Dimension $k \leq n$ kann durch einen zusätzlichen Index gekennzeichnet werden. Nimmt man die Vektoren x_i als Spalten einer Matrix, so erhält man

$$X = X(n,k) = (x_1, x_2, ..., x_k) = [x_1, x_2, ..., x_k].$$

Ein spezielles mittels der Matrix A und einem Vektor $x \neq 0$ generiertes Vektorsystem ist $\{x, Ax, ..., A^m x\}$.
Sind seine ersten k Vektoren linear unabhängig, so bezeichnet man den durch sie aufgespannten Raum als **Krylov-Unterraum**

$$\mathcal{K}_k = \mathcal{K}_k(A, x) = \operatorname{span}\{x, Ax, ..., A^{k-1}x\}. \tag{2.5}$$

Damit ergibt sich die sogenannte **Krylov-Matrix**

$$K^{(k)} = (x, Ax, ..., A^{k-1}x).$$

Ist x ein EV der Matrix A (siehe EWP gemäß Definition 2.21), so erhält man wegen $A^m x = \lambda^m x,\ m \geq 1$, die einfache Situation mit dem Krylov-Unterraum $\mathcal{K}_1 = \mathcal{K}_2 = \mathcal{K}_3 = ... = \text{span}\{x\}$.
Für eine singuläre Matrix macht die Konstruktion von $\mathcal{K}_k$ wenig Sinn.

Beispiel 2.2 Krylov-Unterräume für eine singuläre Matrix A

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

(1) $x = (1,0,0)^T$ EV

$$\mathcal{K}_1 = \text{span}\{x\}, \quad Ax = 0 = 0 \cdot x$$

(2) $x = (0,1,0)^T$ kein EV

$$\mathcal{K}_2 = \text{span}\{x, Ax\}, \quad Ax = (1,0,0)^T,\ A^2x = 0$$

Für reguläre Matrizen ist zu prüfen, wie groß maximal die Dimension von $\mathcal{K}_k$ werden kann.

Beispiel 2.3 Krylov-Unterräume für eine reguläre Matrix A

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

(1) $x = (1,0,0)^T$ EV

$$\mathcal{K}_1 = \text{span}\{x\}, \quad Ax = x = 1 \cdot x$$

(2) $x = (1,1,0)^T$ kein EV

$$\mathcal{K}_2 = \text{span}\{x, Ax\}, \quad Ax = (2,1,0)^T,\ A^m x = (c_1, c_2, 0)^T,\ m \geq 2,\ \mathcal{K}_3 = \mathcal{K}_2$$

(3) $x = (0,0,1)^T$ kein EV

$$\mathcal{K}_3 = \text{span}\{x, Ax, A^2x\}, \quad Ax = (1,1,1)^T,\ A^2x = (3,2,1)^T$$

Die Kontrolle, ob bei Hinzunahme des nächsten Vektors $A^m x$ die Dimension des Vektorraums um Eins wächst, erfolgt in Anwendungen durch die Transformation des Vektorsystems auf eine orthogonale Basis $\{q_1, q_2, ..., q_m, q_{m+1}\}$, $q_i^T q_j = \delta_{ij}$, unter Verwendung z. B. des Orthogonalisierungsverfahrens von Gram-Schmidt. Kann der Vektor q_{m+1} nicht erzeugt werden, ist der maximale Krylov-Unterraum $\mathcal{K}_m(A, x)$.

Die lineare Unabhängigkeit und Orthogonalität von Vektoren sowie die spd-Eigenschaft von Matrizen führen uns auf das Merkmal der Konjugiertheit oder A-Orthogonalität von Vektoren.

Definition 2.6 Konjugiertheit oder A-Orthogonalität von Vektoren

Sei $A = A^T > 0$.

(1) Zwei nicht verschwindende Vektoren $x, y \in \mathbb{R}^n$ *sind konjugiert oder A-orthogonal, wenn ihr Skalarprodukt* $(Ax, y) = x^T Ay$ *Null ist.*

(2) Das System von nicht verschwindenden Vektoren $x_i \in \mathbb{R}^n$ *ist konjugiert oder A-orthogonal, falls* $(Ax_i, x_j) = x_i^T A x_j = 0$ *für alle* $i \neq j$.

Wegen der Symmetrie von A gilt natürlich $(Ax, y) = (Ay, x)$.

Satz 2.1 *Ein System von A-orthogonalen Vektoren ist linear unabhängig.*

Beweis. Seien die Vektoren x_i, $i = 1, 2, ..., k$, A-orthogonal und

$$\sum_{i=1}^{k} c_i x_i = 0, \quad c_i \in \mathbb{R}.$$

Daraus folgt für beliebiges $j \in \{1, 2, ..., k\}$

$$\begin{aligned} 0 &= x_j^T A \sum_{i=1}^{k} c_i x_i \\ &= \sum_{i=1}^{k} c_i x_j^T A x_i \\ &= c_j x_j^T A x_j, \quad x_j^T A x_j > 0, \\ 0 &= c_j, \end{aligned}$$

was die lineare Unabhängigkeit der Vektoren x_i bedeutet. □

Natürlich gilt die Behauptung des Satzes nicht in umgekehrter Richtung.

Wie ist die Situation bei anderen Voraussetzungen über die Matrix A?
Wenn die Matrix zwar symmetrisch, aber indefinit ist, geht uns die Bedingung $x^T Ax > 0$ verloren, die im Beweis ja gebraucht wurde. Wir haben in diesem Fall die A-Orthogonalität eines Vektors zu sich selbst, wie bei

$$x^T Ax = (1, 1) \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0.$$

Anders liegt die Situation bei $A \neq A^T$ und $A > 0$, denn man muss dann i. Allg. mit $x^T Ay \neq y^T Ax$ rechnen.
Als Beispiel nehmen wir die Matrix

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$$

und die Vektoren $x = (1, -1)^T,\ y = (1, 1)^T$. Man erhält

$$x^T Ay = y^T A^T x = 0, \quad y^T Ax = -2.$$

Mit den allgemeinen Ansätzen $(x_1, x_2)A(y_1, y_2)^T = 0$ und $(y_1, y_2)A(x_1, x_2)^T = 0$ lässt sich kein System von reellen A-orthogonalen Vektoren finden.
Zusätzlich gibt es im Beweis zur linearen Unabhängigkeit ein Problem.
Aus $c_1 x + c_2 y = 0$ folgt zwar

$$\begin{aligned} 0 &= c_1 Ax + c_2 Ay \\ &= c_1 \underbrace{x^T Ax}_{>0} + c_2 \underbrace{x^T Ay}_{=0} \\ &= c_1, \end{aligned}$$

aber andererseits

$$\begin{aligned} 0 &= c_1 \underbrace{y^T Ax}_{\neq 0} + c_2 \underbrace{y^T Ay}_{>0} \\ &= c_1 \delta_1 + c_2 \delta_2, \end{aligned}$$

und damit nicht zwingend $c_2 = 0$.

Definition 2.7 Permutationsmatrix
Eine Permutationsmatrix P ist eine quadratische reelle Matrix mit genau einem Einselement je Zeile und Spalte sowie Nullen sonst. Es gilt $PP^T = P^T P = I,\ P^T = P^{-1}$.

Zur vereinfachten Darstellung einer Permutationsmatrix P verwendet man einen Permutationsvektor $p = (p_1, p_2, ..., p_n)$, wobei $(p_1, p_2, ..., p_n)$ ist eine Permutation der natürlichen Zahlen $1, 2, ..., n$ ist. Wird ein solcher im Algorithmus benötigt, dann erfolgt seine Initialisierung zumeist mit $p = (1, 2, ..., n)$. p_i bedeutet in der i-ten Zeile und p_i-ten Spalte von P eine Eins, sonst sind Nullen in der Zeile.
Für $n = 5$ und $p = (4, 3, 5, 1, 2)$ z. B. ist das die Matrix

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Zeilenvertauschungen (ZV) in einer Matrix A bedeuten die Matrixtransformation PA (Linksmultiplikation), Spaltenvertauschungen entsprechen der Rechtsmultiplikation AP.

Definition 2.8 Diagonaldominanz
Die reelle Matrix $A = A(n,n) = (a_{ij})$ heißt streng (strikt) diagonaldominant, wenn

$$|a_{ii}| \; > \; \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, 2, ..., n. \tag{2.6}$$

Gilt die Größerbeziehung mit $\geq$, so spricht man von schwacher Diagonaldominanz.

Definition 2.9 Strenge Regularität
Die Matrix $A \in \mathbb{R}^{n,n}$ ist streng regulär, falls alle Hauptuntermatrizen

$$\widetilde{A}_k = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}, \quad 1 \leq k \leq n, \tag{2.7}$$

regulär sind.

Definition 2.10 Involution
Die Matrix $A \in \mathbb{R}^{n,n}$ heißt involutorisch, falls $A^2 = I$, I Einheitsmatrix.

Involutorische Matrizen treten im Zusammenhang mit dem dyadischen Produkt und Reflexionsmatrizen auf.

Definition 2.11 Asymmetrie
Die Matrix $A = (a_{ij}) \in \mathbb{R}^{n,n}$ heißt asymmetrisch, falls $a_{ij} = -a_{ji}$ für $1 \leq i \neq j \leq n$ gilt.

Definition 2.12 Matrix irreduzibel diagonaldominant
Die reelle Matrix $A = A(n,n) = (a_{ij})$ heißt irreduzibel diagonaldominant, wenn
(1)

$$|a_{ii}| \; \geq \; \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, 2, ..., n, \tag{2.8}$$

und für mindestens ein i gilt die strenge Größerbeziehung,

(2) und es gibt keine Permutationsmatrix P, mit der eine Transformation der Matrix gemäß

$$\widetilde{A} \; = \; PAP^T \; = \; \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{pmatrix}, \quad \widetilde{A}_{11} \in \mathbb{R}^{p,p}, \; \widetilde{A}_{22} \in \mathbb{R}^{q,q}, \; p+q = n, \tag{2.9}$$

möglich ist. Das bedeutet die Irreduzibilität der Matrix.
Falls es eine solche Transformation gibt, ist die Matrix reduzibel.

Entscheidend für die Irreduzibilität einer Matrix ist die Lage ihrer NNE. Dabei könnten wir die NNE der Matrix A einfach zu Eins definieren.
Ein graphentheoretischer Zugang erleichert die Überprüfung dieser Eigenschaft.
Zur Matrix $A = A(n,n) = (a_{ij})$ konstruiert man den Graphen $G(A)$ mit den Knoten $P_1, P_2, ..., P_n$ und den gerichteten Kanten $P_i \to P_j$, die entstehen gdw. $a_{ij} \neq 0$ ist. Hat man andererseits den Graphen mit seinen Knoten und Kanten, so ist die Adjazenzmatrix A_{adj} definiert. Diese hat genau an der Stelle (i,j) eine Eins stehen, wenn eine gerichtete Kante $P_i \to P_j$ auftritt, sonst hat sie Nulleinträge.

Beispiel 2.4 Matrix, ihr Graph und Adjazenzmatrix
(1)

$$A(3,3) = \begin{pmatrix} 1 & 2 & 0 \\ -1 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix},$$

Graph $G(A)$: $\{P_1, P_2, P_3;\ P_1 \to P_1, P_1 \leftrightarrow P_2, P_2 \to P_2, P_3 \to P_3, P_3 \to P_1\}$,

$$\text{Adjazenzmatrix } A_{adj} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

(2)

$$A(n,n) = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix},$$

Graph $G(A)$: $\{P_1, P_2, ..., P_n;$
$P_1 \to P_1, P_1 \leftrightarrow P_2, P_2 \to P_2, P_2 \leftrightarrow P_3, ..., P_{n-1} \leftrightarrow P_n, P_n \to P_n\}$,

$$\text{Adjazenzmatrix } A_{adj} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{pmatrix}.$$

Die Matrix A ist genau dann irreduzibel, wenn ihr Graph $G(A)$ stark zusammenhängend ist, d. h. für jedes Knotenpaar $(P_i, P_j),\ 1 \leq i,j \leq n$, gibt es einen gerichteten Weg $P_i \to P_{k_1} \to ... \to P_{k_m} \to P_j$ von P_i nach P_j.

Somit ist im Beispiel 2.4 die Matrix $A(3,3)$ reduzibel, weil das Paar (P_2, P_3) nicht über einen gerichteten Weg zu verbinden ist.
Mit der Permutationsmatrix P auf der Basis von $p = (3,2,1)$ für die Matrixtransformation erhält man

$$\widetilde{A} \;=\; PAP^T \;=\; \left(\begin{array}{c|cc} 1 & 0 & 3 \\ \hline 0 & 1 & -1 \\ 0 & 2 & 1 \end{array}\right) \;=\; \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{pmatrix}.$$

Die Tridiagonalmatrix $A(n,n)$ ist hingegen irreduzibel, weil man von jedem Knoten zu jedem anderen über einen gerichteten Weg gelangen kann.
Allgemein besitzen Matrizen, die aus Diskretisierungsmethoden für RWA entstehen und eine tridiagonale, machmal noch eine symmetrische Struktur bzw. Blockstruktur haben, diese Eigenschaft.

Die Reduzibilität einer Matrix, wie sie in der Definition 2.12 vorgestellt worden ist, lässt sich auf eine andere äquivalente Art und Weise anschaulich erläutern. Dabei korrespondiert die Blockzerlegung der Matrix mit der Bildung von zwei Indexmengen auf der Basis von Nullelementen der Matrix.

Definition 2.13 Matrix reduzibel
Die reelle Matrix $A = A(n,n) = (a_{ij})$ heißt reduzibel, wenn zwei Indexmengen $S_1, S_2 \subset W = \{1,2,...,n\}$ mit den Bedingungen

$$S_1 \neq \emptyset,\; S_2 \neq \emptyset, \quad S_1 \cup S_2 = W,\; S_1 \cap S_2 = \emptyset,$$

($\emptyset$ ist die leere Menge) existieren, so dass $a_{ij} = 0$ für alle $i \in S_1$ und $j \in S_2$ ist.

Beispiel 2.5 Reduzible Matrizen
(1) Matrix aus Beispiel 2.4

$$A(3,3) = \begin{pmatrix} 1 & 2 & 0 \\ -1 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}.$$

Wir haben $S_1, S_2 \subset W = \{1,2,3\}$ und mit $a_{13} = 0,\; a_{23} = 0$ die Indexmengen $S_1 = \{1,2\},\; S_2 = \{3\}$. Das Element $a_{32} = 0$ kann also keine Berücksichtigung finden.

(2) Matrix mit symmetrisch liegenden Nullelementen

$$A(3,3) = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Wir haben $S_1, S_2 \subset W = \{1,2,3\}$ und mit $a_{13} = a_{31} = 0,\; a_{23} = a_{32} = 0$ kann man sowohl die Indexmengen $S_1 = \{1,2\},\; S_2 = \{3\}$ also auch die Mengen $S_1 = \{3\},\; S_2 = \{1,2\}$ wählen.

(3) Matrix mit symmetrisch liegenden NNE (als $*$ eingetragen)

$$A(6,6) = \begin{pmatrix} * & 0 & * & 0 & 0 & * \\ 0 & * & 0 & * & * & 0 \\ * & 0 & * & 0 & 0 & * \\ 0 & * & 0 & * & * & 0 \\ 0 & * & 0 & * & * & 0 \\ * & 0 & * & 0 & 0 & * \end{pmatrix}.$$

Wir können hier $S_1, S_2 \subset W = \{1, 2, ..., 6\}$ als $S_1 = \{1, 3, 6\},\ S_2 = \{2, 4, 5\}$ wählen. Um die Mischung der Indexmengen aufzuheben, brauchen wir nur die Werte 6 und 2 auszutauschen. Dies ergibt den Permutationsvektor $p = (1, 6, 3, 4, 5, 2)$, die zugehörige Permutationsmatrix P sowie die Transformationsbeziehung

$$PAP^T = \begin{pmatrix} * & * & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * & * \end{pmatrix} = \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{pmatrix}.$$

Für die Lösung eine LGS $Ax = b$ bedeutet die Reduzibilität der Koeffizientenmatrix, dass das LGS nach Umordnen in kleinere Systeme zerlegt werden kann, die nacheinander zu lösen sind.

Manchmal finden wir in den Matrixelementen eine bestimmte Vorzeichensituation vor, denken wir nur an die FDM für RWA. Es sind also ganz spezielle Matrizen.

Definition 2.14 L-Matrix
Die reelle Matrix $A = (a_{ij})$ heißt L*-Matrix, falls*

$$a_{ii} > 0,\ i = 1, 2, ..., n, \quad \textit{und} \quad a_{ij} \leq 0 \textit{ für } i \neq j.$$

Definition 2.15 Stieltjes-Matrix
Die reelle Matrix A ist eine Stieltjes-Matrix, falls sie L*-Matrix und $A = A^T > 0$ ist.*

Da aus $A = A^T > 0$ die Beziehung $a_{ii} > 0$ folgt, kann man die Stieltjes-Matrix auch als spd Matrix mit nicht positiven Nichtdiagonalelementen definieren.

Definition 2.16 M-Matrix
Wir erwähnen von den zahlreich existierenden äquivalenten Definitionen zwei.
Die reelle Matrix A ist eine M*-Matrix, falls gilt:*
(1.1) $a_{ij} \leq 0$ *für* $i \neq j$,
(1.2) $\exists\ A^{-1} = (a'_{ij})$ *mit* $\forall\ i, j\ \ a'_{ij} \geq 0$,
oder
(2.1) $a_{ij} \leq 0$ *für* $i \neq j$,
(2.2) $\exists\ x \in \mathbb{R}^n$ *mit* $x_i > 0$ *und* $(Ax)_i > 0$ *für alle* $i = 1, 2, ..., n$.

Definition 2.17 Monotone Matrix
Die reelle Matrix A ist monoton oder invers-isoton, falls gilt:
(1) $\exists\ A^{-1} = (a'_{ij})$,
(2) $\forall\ i,j\quad a'_{ij} \geq 0$.

Definition 2.18 *Sei* $x, y \in \mathbb{R}^n$. *Die reelle Matrix A ist:*
(1) isoton, wenn komponentenweise aus $x \leq y$ *die Ungleichung* $Ax \leq Ay$ *folgt,*
(2) antiton, wenn komponentenweise aus $x \leq y$ *die Ungleichung* $Ay \leq Ax$ *folgt,*
(3) monoton oder invers-isoton, wenn komponentenweise aus $Ax \leq Ay$ *die Ungleichung* $x \leq y$, *d. h. auch* $0 \leq Ax \ \Rightarrow\ 0 \leq x$, *folgt.*

Beispiel 2.6 Einige Eigenschaften von L-Matrizen

A	A^{-1}	Eigenschaften
$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$	$\not\exists$	• schwach diagonaldominant, $A = A^T \geq 0$, keine M-Matrix.
$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$	$\frac{1}{3}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$	• streng diagonaldominant, irreduzibel diagonaldominant, $A = A^T > 0$, Stieltjes-Matrix, M-Matrix, monotone Matrix.
$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$	• schwach diagonaldominant, irreduzibel diagonaldominant, $A = A^T > 0$, Stieltjes-Matrix, M-Matrix, monotone Matrix.
$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$	$\frac{1}{4}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}$	• schwach diagonaldominant, irreduzibel diagonaldominant, $A = A^T > 0$, Stieltjes-Matrix, M-Matrix, monotone Matrix.

Definition 2.19 Matrix mit Eigenschaft A
Die reelle Matrix $A(n,n)$ *hat die Eigenschaft* A, *falls eine Permutationsmatrix P existiert, so dass*

$$PAP^T = \begin{pmatrix} D_1 & M_1 \\ M_2 & D_2 \end{pmatrix}, \tag{2.10}$$

wobei $D_1(p,p),\ D_2(q,q),\ p+q=n$, *(nicht unbedingt reguläre) Diagonalmatrizen sind.*

Da durch Zeilen- und Spaltenvertauschungen die Matrixelemente wertemäßig nicht verändert werden, hat eine solche Matrix A viele Nulleinträge, was z. B. auch bei Bandmatrizen der Fall ist. Außerdem ist $PAP^T = PAP^{-1}$ ähnlich zu A.

Beispiel 2.7 Matrizen mit der Eigenschaft A

(1)

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad PAP^T = \left(\begin{array}{c|c} D_1 & M_1 \\ \hline M_2 & D_2 \end{array}\right), \quad P = I,\ D_1 = D_2 = (1).$$

(2)

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 3 & 0 & 4 \\ 0 & 2 & 2 \end{pmatrix},$$

$$PAP^T = \left(\begin{array}{c|c} D_1 & M_1 \\ \hline M_2 & D_2 \end{array}\right) = \left(\begin{array}{c|cc} 0 & 3 & 4 \\ \hline -1 & 1 & 0 \\ -2 & 0 & 2 \end{array}\right),$$

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad D_1 = (0),\ D_2 = \mathrm{diag}(1, 2).$$

Bei der direkten Lösung von LGS $Ax = b$ ist ein wichtiger Aspekt die Überführung der Matrix (meist mit rechter Seite) in eine obere Dreiecksform. Dies entspricht dem GA in seiner Originalform.

Die entscheidende Idee dabei ist demnach:

Systematische Erzeugung von Nullen in der Matrix durch Subtraktion des Vielfachen einer Zeile von einer anderen.

Definition 2.20 Allgemeine Reduktion der Matrix $A = (a_{ij})$

Die Reduktion wird auf dem Platz von $A^{(0)} = A,\ a_{ij}^{(0)} = a_{ij}$, *vorgenommen.*

Die Durchführbarkeit wird vorausgesetzt, d. h. $a_{kk}^{(k-1)} \neq 0$.

Schritte

S1: $k = 1,\ a_{11}^{(0)} \neq 0$, *Spaltenelemente* $a_{21}^{(0)}, a_{31}^{(0)}, ..., a_{n1}^{(0)}$ *zu Null machen.*

Weg: i-te Zeile := i*-te Zeile* $-\dfrac{a_{i1}^{(0)}}{a_{11}^{(0)}} * 1.$ *Zeile,* $i = 2, 3, ..., n,$

bzw. $a_{ij}^{(1)} = a_{ij}^{(0)} - \dfrac{a_{i1}^{(0)} a_{1j}^{(0)}}{a_{11}^{(0)}},\ i,j = 2,3,...,n.$

Ergebnis mit $A^{(1)}$*:*

$$\begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3n}^{(1)} \\ . & \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

S2: Resttableau $A^{(1)}$ *der Dimension* $n-1$ *analog verarbeiten.*

$k = 2,\ a_{22}^{(1)} \neq 0$ *Spaltenelemente* $a_{32}^{(1)}, a_{42}^{(1)}, ..., a_{n2}^{(1)}$ *zu Null machen.*

Weg: i-te Zeile := i*-te Zeile* $-\dfrac{a_{i2}^{(1)}}{a_{22}^{(1)}} *$ *2. Zeile,* $i = 3,4,...,n.$

Ergebnis mit $A^{(2)}$*:*

$$\begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ . & . & \cdots & \cdots & \cdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}.$$

S3: Resttableau $A^{(2)}$ *der Dimension* $n-2$ *analog verarbeiten usw.*

Sn-1: Endtableau der Reduktion

$$\begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & a_{nn}^{(n-1)} \end{pmatrix}.$$

Das jeweils entstehende Resttableau ist die k*-te reduzierte Matrix*

$$A^{(k)} = \begin{pmatrix} a_{k+1,k+1}^{(k)} & a_{k+1,k+2}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ a_{k+2,k+1}^{(k)} & a_{k+2,k+2}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n,k+1}^{(k)} & a_{n,k+2}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix},\ \textit{wobei} \tag{2.11}$$

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)} a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}},\ i,j = k+1, k+2, ..., n;\ k = 1,2,...,n-1. \tag{2.12}$$

Natürlich kann man wegen

$$a_{ij}^{(k)} \;=\; a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)} a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}} \;=\; a_{ij}^{(k-1)} + \Big(- \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \Big)\, a_{kj}^{(k-1)}$$

von der Addition des negativen Vielfachen einer Zeile zu einer anderen sprechen.

Beispiel 2.8 Gauß-Reduktion, GA
Wir demonstrieren den Algorithmus im Zusammenhang mit der Lösung des LGS $Ax = b$, d. h. wir behandeln auf analoge Weise zusätzlich die rechte Seite.

$A^{(0)}$

1	0	3	2	6
4	1	8	3	16
0	1	-3	-4	-6
2	0	8	5	15

Vielfache zur 1. Spalte: 4, 0, 2

$A^{(1)}$

1	0	3	2	6
	1	-4	-5	-8
	1	-3	-4	-6
	0	2	1	3

Vielfache zur 2. Spalte: 1, 0

$A^{(2)}$

1	0	3	2	6
	1	-4	-5	-8
		1	1	2
		2	1	3

Vielfache zur 3. Spalte: 2

$A^{(3)}$

1	0	3	2	6
	1	-4	-5	-8
		1	1	2
			-1	-1

Die Lösung des transformierten LGS $A^{(3)}x = c$ erfolgt durch Rückwärtseinsetzen, damit sind $x_4 = 1,\ x_3 = 1,\ x_2 = 1,\ x_1 = 1$.

Natürlich ergeben sich sofort einige Fragen, auf die später genauer eingegangen wird. Die Matrix $A = A^{(0)}$ wird durch eine obere Dreiecksmatrix $A^{(n-1)} = U$ überschrieben, an den entstehenden "Nullpositionen" könnte man sich spaltenweise die Vielfachen merken, und zusammen mit der Eins-Diagonalen ergibt dies eine untere Dreiecksmatrix L. Im Ergebnis bekommen wir die Faktorisierungkomponenten L ($l_{ii} = 1$), U von A, also $A = LU$.
Strategien gibt es auch für den Fall $a_{kk}^{(k-1)} = 0$, die auf eine Pivotisierung führen.

Weiterhin betrachten wir das Matrixeigenwertproblem (EWP).

Definition 2.21 Eigenwertproblem

(1) **Allgemeines EWP**

Sei $A = (a_{ij})$ eine reelle oder komplexe (n, n)-Matrix.
$\lambda \in \mathbb{C}$ heißt Eigenwert (EW) oder charakteristischer Wert von A, falls ein Vektor $x \in \mathbb{C}^n,\ x \neq 0$, existiert, so dass

$$Ax \ = \ \lambda x \tag{2.13}$$

gilt. x heißt (zugehöriger) Eigenvektor (EV), Rechts-EV oder Eigenlösung.

Zuweilen wird auch das Links-EWP $y^T A = \lambda y^T,\ y \neq 0$, betrachtet.
y ist Links-EV genau dann, wenn y EV von A^T ist.
Ein EV ist bis auf eine multiplikative Konstante eindeutig definiert. So kann er auf ganz unterschiedliche Weise skaliert oder normiert werden.

(2) **Lineares oder klassisches EWP**

$$Ax \ = \ \lambda x \quad mit \quad A \in \mathbb{R}^{n,n},\ \lambda \in \mathbb{C},\ x \in \mathbb{C}^n,\ x \neq 0.$$

(3) **Verallgemeinertes lineares EWP**

$$Ax \ = \ \lambda Bx \quad mit \quad A, B \in \mathbb{C}^{n,n},\ \lambda \in \mathbb{C},\ x \in \mathbb{C}^n,\ x \neq 0.$$

(4) **Nichtlineares EWP**

$$A(\lambda)x \ = \ 0 \quad mit \quad A(\lambda) \in \mathbb{C}^{n,n},\ \lambda \in \mathbb{C},\ x \in \mathbb{C}^n,\ x \neq 0.$$

(5) **Darstellung des EWP als homogenes LGS mit singulärer Koeffizientenmatrix**

$$(A - \lambda I)x \ = \ 0 \quad bzw. \quad (A - \lambda B)x \ = \ 0, \quad I \ Einheitsmatrix.$$

Die Matrix $A_\lambda = A - \lambda I$ ist singulär, also $\det(A_\lambda) = 0$, und hat somit den Rang kleiner als n (Anzahl ihrer linear unabhängigen Zeilen oder Spalten). Mit dem Rang $\mathrm{rang}(A_\lambda)$ ist auch der Rangabfall $n - \mathrm{rang}(A_\lambda)$ definiert.

(6) **Charakteristisches Polynom und charakteristische Gleichung zum EWP**

$$\begin{aligned}
p_n(\lambda) \ &= \ \det(A - \lambda I) \ = \ c_0\lambda^n + c_1\lambda^{n-1} + ... + c_{n-1}\lambda + c_n, \quad c_0 = (-1)^n, \\
p_n(\lambda) \ &= \ 0, \quad \mathrm{Grad}(p_n) = n. \\
q_n(\lambda) \ &= \ \det(A - \lambda B) \ = \ \tilde{c}_0\lambda^n + \tilde{c}_1\lambda^{n-1} + ... + \tilde{c}_{n-1}\lambda + \tilde{c}_n, \\
q_n(\lambda) \ &= \ 0, \quad \mathrm{Grad}(q_n) \leq n.
\end{aligned} \tag{2.14}$$

Beispiel 2.9 EWP zur Matrix $A = A^T > 0$

$$A(3,3) = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix},$$

charakteristisches Polynom $p_3(\lambda) = \lambda^3 - 6\lambda^2 + 10\lambda - 4$,
paarweise verschiedene EW mit orthogonalen EV

$\lambda_1 = 2 - \sqrt{2}, \quad v_1 = (1, \sqrt{2}, 1)^T$
$\lambda_2 = 2, \qquad\quad v_2 = (1, 0, -1)^T$
$\lambda_3 = 2 + \sqrt{2}, \quad v_3 = (1, -\sqrt{2}, 1)^T$

Beispiel 2.10 EWP
Für eine einfache indefinite Matrix führen wir einige Rechnungen in Maple aus.
Definition der Matrix numerisch und symbolisch

```
> restart;
  with(linalg):

> Digits:=20:
> n:=2:
  A:=matrix(n,n,[[0.780,0.563],
                 [0.913,0.659]]);
  Ar:=matrix(n,n,[[39/50,563/1000],
                  [913/1000,659/1000]]);
```

$$n := 2$$

$$A := \begin{bmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{bmatrix}$$

$$Ar := \begin{bmatrix} \frac{39}{50} & \frac{563}{1000} \\ \frac{913}{1000} & \frac{659}{1000} \end{bmatrix}$$

```
From the Help and Description:  Eigenvals(A)
```

- `Eigenvals(A) returns an array of all eigenvalues of A.`
  `The eigenvalues are computed by the QR method.`
  `The matrix is first balanced and transformed into upper Hessenberg form.`
  `Then the eigenvalues (eigenvectors) are computed.`
- `If the matrix is symmetric then the routine will handle the matrix specially (using a faster algorithm).`
- `Note that all the entries of A must be numerical.`
- `The function Eigenvals itself is inert. To actually compute the eigenvalues and eigenvectors, the user must evaluate the inert function in the floating point domain, by evalf(Eigenvals(A)).`

```
> II:=array(identity,1..n,1..n):
  lambda:='lambda':
  evalm(A-lambda*II);
  expand(det(A-lambda*II),lambda);
  charpoly(Ar,lambda);
  solve(det(A-lambda*II),lambda);

  vecsa:='vecsa':
  lambda:=evalf(Eigenvals(A,vecsa));
  EV:=evalm(vecsa);   # EV nicht normiert
```

$$\begin{bmatrix} 0.780-\lambda & 0.563 \\ 0.913 & 0.659-\lambda \end{bmatrix}$$

$$0.1\,10^{-5}-1.439\,\lambda+\lambda^2$$

$$\frac{1}{1000000}-\frac{1439}{1000}\lambda+\lambda^2$$

$$1.4389993050726317415,\ 0.69492736825854562526\,10^{-6}$$

$$\lambda := [1.4389993050726317415,\ 0.694927368258 54\,10^{-6}]$$

$$EV := \begin{bmatrix} 0.64955572831439685711 & -0.60232599023662408990 \\ 0.76031398501800126831 & 0.83448286645071325268 \end{bmatrix}$$

```
> EV1:=col(EV,1);
  EV2:=col(EV,2);
```

$$EV1 := [0.64955572831439685711,\ 0.76031398501800126831]$$

$$EV2 := [-0.60232599023662408990,\ 0.83448286645071325268]$$

Kontrolle der Länge der EV

```
> EV1[1]^2+EV1[2]^2;
  EV2[1]^2+EV2[2]^2;
```

$$1.0000000000000000000$$

$$1.0591582529143287084$$

```
> la_ex:=solve(det(Ar-la_ex*II),la_ex);
```

$$la\_ex := \frac{1439}{2000}+\frac{\sqrt{2070717}}{2000},\ \frac{1439}{2000}-\frac{\sqrt{2070717}}{2000}$$

From the Help and Description: eigenvalues(A)

- The call eigenvalues(A) or eigenvals(A) returns a sequence of the eigenvalues of A.
- All the entries of A must be numerical or symbolic.
- If A contains only numerical entries and at least one floating-point number, a numerical method is used where all arithmetic is done at the precision specified by Digits.
  In this case eigenvalues(A) returns a sequence of all eigenvalues of A.
- Otherwise (no floating-point numbers, i.e. the symbolic case), the eigenvalues are computed by solving the characteristic polynomial det(lambda*I-A)=0 for the scalar variable lambda, where I is the identity matrix.

- In the symbolic case (i.e an entry is an unassigned name), the eigenvalues are expressed using Maple's RootOf notation for algebraic extensions. Maple tries to express the eigenvalues in terms of exact radicals. Note that if the characteristic polynomial has a factor of degree greater than four, then it may not be possible to express all the eigenvalues in terms of radicals.
- Problem in Maple V:
  By the matrix dimension is greater then 3.
  If the matrix entries are more complicated (i.e. see Hilbert matrix), then eigenvals(A) returns only one of the eigenvalues of A.
  In this case the characteristic polynomial has a factor of degree greater than 3, then it may not be possible to express all the eigenvalues in terms of radicals.
  Then one does:

```
allvalues(eigenvals(A)):        # RootOf - notation
evalf(allvalues(eigenvals(A))); # evaluation
```

```
> la:=eigenvalues(A);
  lar:=eigenvalues(Ar);
```

$$la := 1.4389993050726317415,\ 0.694927368258541\,10^{-6}$$

$$lar := \frac{1439}{2000} + \frac{\sqrt{2070717}}{2000},\ \frac{1439}{2000} - \frac{\sqrt{2070717}}{2000}$$

From the Help an Description: eigenvectors(A)

- The procedure eigenvectors(A) or eigenvects(A) computes the eigenvalues and eigenvectors of A. That is, for each eigenvalue lambda of A it solves the linear system (lambda*I-A)x=0 for x.
- The result returned is a sequence of lists of the form [ei, mi, {v[1,i],... v[ni,i]}], where the ei are the eigenvalues, mi their algebraic multiplicities, {v[1,i], ..., v[ni,i]} is a set of basis vectors for the eigenspace corresponding to ei, and 1 <= ni <= mi is the dimension of the eigenspace.
- Numeric Case: If the matrix A contains any floating-point (decimal) numbers, the eigenvectors are computed numerically.
  A standard numerical algorithm is used. All floating point arithmetic is done at Digits digits of precision. Note that the matrix entries on input must all be all of type numeric or complex(numeric).
- Symbolic Case: Otherwise the eigenvalues and eigenvectors are computed symbolically (exactly). First the characteristic polynomial is computed and solved for it's roots (the eigenvalues) lambda[i] symbolically. Then for each eigenvalue, a basis for it's eigenspace is computed by computing the null space of the characteristic matrix A-lambda[i]*I.

Achtung: EV als Ergebnis der symbolischen Rechnung können ganz unterschiedliche Skalierung haben.

```
> eigenvectors(A);     # EV i. Allg. nicht normiert, aber skaliert
  ewv1:=eigenvectors(Ar);
```

$$[1.4389993050726317414,\ 1,\ \{[0.64955572831439685712,\ 0.76031398501800126831]\}],$$
$$[0.69492736825854\,10^{-6},\ 1,\ \{[-0.60232599023662408990,\ 0.83448286645071325268]\}]$$

$$ewv1 := \left[\frac{1439}{2000}+\frac{\sqrt{2070717}}{2000},\ 1,\ \left\{\left[1,\ -\frac{121}{1126}+\frac{\sqrt{2070717}}{1126}\right]\right\}\right],$$
$$\left[\frac{1439}{2000}-\frac{\sqrt{2070717}}{2000},\ 1,\ \left\{\left[1,\ -\frac{121}{1126}-\frac{\sqrt{2070717}}{1126}\right]\right\}\right]$$

```
> v11:=evalf(op(ewv1[1][3]));
  v12:=evalf(op(ewv1[2][3]));
```

$$v11 := [1.,\ 1.1705138633616904822]$$
$$v12 := [1.,\ -1.3854339344096478534]$$

Normierung mit der euklidischen Norm

```
> no1:=sqrt(v11[1]^2+v11[2]^2):
  no2:=sqrt(v12[1]^2+v12[2]^2):
  evalm(v11/no1);
  evalm(v12/no2);
```

$$[0.64955572831439685710,\ 0.76031398501800126831]$$
$$[0.58526314402966098506,\ -0.81084342029797362184]$$

```
> ewv2:=eigenvectors(Ar);  # andere Skalierung der EV
```

$$ewv2 := \left[\frac{1439}{2000}+\frac{\sqrt{2070717}}{2000},\ 1,\ \left\{\left[\frac{11}{166}+\frac{\sqrt{2070717}}{1826},\ 1\right]\right\}\right],$$
$$\left[\frac{1439}{2000}-\frac{\sqrt{2070717}}{2000},\ 1,\ \left\{\left[\frac{11}{166}-\frac{\sqrt{2070717}}{1826},\ 1\right]\right\}\right]$$

```
> v21:=evalf(op(ewv2[1][3]));
  v22:=evalf(op(ewv2[2][3]));
```

$$v21 := [0.85432563534789895011,\ 1.]$$
$$v22 := [-0.72179551486597123927,\ 1.]$$

```
> no1:=sqrt(v21[1]^2+v21[2]^2):
  no2:=sqrt(v22[1]^2+v22[2]^2):
  evalm(v21/no1);
  evalm(v22/no2);
```

$$[0.64955572831439685713,\ 0.76031398501800126832]$$
$$[-0.58526314402966098506,\ 0.81084342029797362183]$$

Kontrolle von $Ax = \lambda x$

```
> evalf(evalm(A&*v21-la[1]*v21));
```

$$[0.1\,10^{-18},\ 0.]$$

```
> evalf(evalm(A&*v22-la[2]*v22));
```

$$[-0.272152\,10^{-20},\ 0.1\,10^{-19}]$$

2.1.1 Matrixzerlegungen

Betrachten wir die reelle Matrix $A = A(n,n) = (a_{ij})$.
Eine beliebige Zerlegung dieser Matrix kann man in der Form

$$A \;=\; A_1 + A_2 + ... + A_k$$

notieren. Für eine Weiterverarbeitung in numerischen Algorithmen sind jedoch nur spezielle Zerlegungen von Interesse.

Definition 2.22 Reguläre Zerlegung
Unter einer regulären Zerlegung (Splitting) von A versteht man die Darstellung gemäß

$$A \;=\; N - P, \quad \textit{wobei} \quad \det(N) \neq 0. \tag{2.15}$$

Definition 2.23 Spezielle Zerlegung
Eine spezielle Zerlegung der Matrix A mit $a_{ii} \neq 0$ ist

$$A \;=\; D - E - F \;=\; D - C \;=\; D(I - L - U) \tag{2.16}$$

mit
I *Einheitsmatrix,*
D *Diagonalmatrix,* $D = \mathrm{diag}(A) = \mathrm{diag}(a_{11}, a_{22}, ..., a_{nn})$ *regulär,*
L *linke Dreiecksmatrix (auch Diagonale ist Null),* $L = (l_{ij})$, $E = DL$,
U *rechte Dreiecksmatrix (auch Diagonale ist Null),* $U = (u_{ij})$, $F = DU$,
$C = D - A = E + F$ *mit den negativen Werten der Nichtdiagonalelem. von A.*

Natürlich kann man auch eine der Dreiecksmatrizen jeweils mit dem Diagonalanteil zusammenfassen oder den Diagonalanteil halbieren wie in

$$A = (D - E) - F = (D - F) - E = (\tfrac{1}{2}D - E) + (\tfrac{1}{2}D - F).$$

Gelegentlich braucht man die spalten- oder zeilenweise Darstellung der Matrix.
Dann schreibt man

$$A \;=\; (a_1, a_2, ..., a_n) = \begin{pmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{pmatrix} \quad \text{oder} \quad A \;=\; \begin{pmatrix} — & a_1 & — \\ — & a_2 & — \\ & \vdots & \\ — & a_n & — \end{pmatrix},$$

wobei man deutlich sagen sollte, ob mit a_i Matrixspalten bzw. -zeilen gemeint sind.
Die Spaltenversion lässt sich auch mit den Einheitsvektoren e_i (Spaltenvektor) als

$$A \;=\; \sum_{i=1}^{n} a_i e_i^T$$

darstellen, wobei $a_i e_i^T$ ein dyadisches Produkt ist.
Viele Matrixzerlegungen beinhalten eine Blockstrukturierung, wie wir sie bei den einführenden Beispielen in Kapitel 1 gesehen haben.
Bei komplexen Matrizen haben wir die Zerlegung in Real- und Imaginärteil.

2.1.2 Eigenschaften von Matrizen

Nun werden wir eine Reihe von wichtigen Eigenschaften von Matrizen als Sätze formulieren und meistens auch beweisen.

Satz 2.2 *Sei P eine Permutationsmatrix.*
Dann gilt $P^{-1} = P^T$ bzw. $PP^T = P^TP = I$, I Einheitsmatrix.

Eine Permutationsmatrix ist also orthogonal.

Satz 2.3 *Wenn die Matrix A hermitesch ist, dann sind alle ihre EW $\lambda(A)$ reell. Insbesondere trifft das auf eine reelle symmetrische Matrix A zu.*

Beweis. (Skizze) Seien $A = A^H = \bar{A}^T$ und $A^* = A^H$ die adjungierte Matrix. Die Verwendung des Skalarproduktes in $\mathbb{C}^n$ und seiner Eigenschaften ergibt

$$\begin{aligned}
(x, Ax) &= (x, \lambda x) = \lambda(x, x), \quad x \neq 0, \\
(x, Ax) &= (A^*x, x) = (A^Hx, x) = (Ax, x) = (\lambda x, x) = \bar{\lambda}(x, x), \\
0 &= (\bar{\lambda} - \lambda)(x, x), \\
\lambda &= \bar{\lambda}.
\end{aligned}$$

Die EV von $A = A^T$ als nicht triviale Lösungen des reellen LGS $(A - \lambda I)x = 0$ sind reelle Vektoren. □

Satz 2.4 *Die Matrix A ist genau dann hermitesch, wenn die quadratische Form $(v, Av) = v^HAv$ für alle $v \in \mathbb{C}^n$ reell ist.*

Beweis.
1. $\Rightarrow$:
Sei $A = A^H$. Die Verwendung des Skalarproduktes von Vektoren in $\mathbb{C}^n$ und seiner Eigenschaften ergibt

$$(v, Av) = (v, A^Hv) = (Av, v) = \overline{(v, Av)}$$

und somit, dass die quadratische Form (v, Av) reell ist.

2. $\Leftarrow$:
Sei $(v, Av) = v^HAv$ reell für alle $v \in \mathbb{C}^n$ und $A = \Re(A) + \imath\Im(A) = B + \imath C$ eine Zerlegung der Matrix in ihren reellen Real- und Imaginärteil.
Wir führen eine Fallunterscheidung und schrittweisen Nachweis von $A^H = B^T - \imath C^T = B + \imath C = A$ durch und zeigen zuerst $C = -C^T$, dann $B = B^T$.

(a) Sei v reell. Daraus folgen die Beziehungen

$$\begin{aligned}(v,Av) &= (v,Bv)+\imath(v,Cv) \quad \Rightarrow \quad (v,Cv)=0,\\ (v,C^Hv) &= (v,C^Tv)=\overline{(C^Tv,v)}=\overline{(v,Cv)}=0,\\ (v,Cv)+(v,C^Tv) &= (v,(C+C^T)v)=(v,\widetilde{C}v)=0, \quad \widetilde{C}^T=\widetilde{C}.\end{aligned}$$

Sei v der k-te Einheitsvektor. Damit erhalten wir

$$(v,(C+C^T)v) \;=\; \tilde{c}_{kk} \;=\; 2c_{kk} \;=\; 0 \quad \Rightarrow \quad \tilde{c}_{kk}=0,\; k=1,2,...,n.$$

Sei $v=(0,...,0,1,0,...,0,1,0,...,0)^T$, also ein Vektor mit Nullkomponenten und genau zwei Einskomponenten an den Stellen k,j. Dann ist

$$(v,(C+C^T)v) \;=\; \tilde{c}_{kk}+\tilde{c}_{kj}+\tilde{c}_{jk}+\tilde{c}_{jj} \;=\; 2\tilde{c}_{kj} \;=\; 0 \quad \Rightarrow \quad \tilde{c}_{kj}=0,\; k,j=1,2,...,n.$$

Alle Elemente von $\widetilde{C}$ sind somit Null, d. h. $C=-C^T$.

(b) Sei v komplex und $v=u+\imath w$.

Folgende Betrachtungen führen auf die zweite Bedingung $B=B^T$.

$$\begin{aligned}(v,Av) &= (u+\imath w,(B+\imath C)(u+iw))\\ &= (u,Bu-Cw)+(w,Bw+Cu)+\\ &\qquad \imath[-(w,Bu-Cw)+(u,Bw+Cu)] \;\in\; \mathbb{R}.\end{aligned}$$

Damit verschwindet der Imaginärteil.

$$\begin{aligned}0 &= -(w,Bu-Cw)+(u,Bw+Cu)\\ &= -(w,Bu)+(w,Cw)+(u,Bw)+(u,Cu),\\ &\qquad (x,Cx)=0 \;\text{ für }\; x\in\mathbb{R}^n \;\text{ wegen (a)}\\ &= (u,Bw)-(w,Bu)\\ &= (u,Bw)-(u,B^Tw)\\ &= (u,(B-B^T)w)\\ &= (u,\widetilde{B}w), \quad \widetilde{B}=B-B^T.\end{aligned}$$

Mit u als k-ten und w als j-ten Einheitsvektor erhalten wir für die Elemente der Matrix $\widetilde{B}$ die Beziehung $\tilde{b}_{kj}=0,\; k,j=1,2,...,n$.
Damit folgt $B=B^T$. □

Satz 2.5 *Ist die Matrix $A \in \mathbb{C}^{n,n}$ positiv definit und hermitesch, dann gilt dies auch für A^{-1} und alle Hauptuntermatrizen*

$$\tilde{A} = \begin{pmatrix} a_{kk} & a_{k,k+1} & \cdots & a_{km} \\ a_{k+1,k} & a_{k+1,k+1} & \cdots & a_{k+1,m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{mk} & a_{m,k+1} & \cdots & a_{mm} \end{pmatrix}, \quad 1 \le k \le m \le n. \tag{2.17}$$

Weiter sind $\det(A) > 0$ *und alle Determinanten von Hauptuntermatrizen (Hauptminoren)* $\det(\tilde{A})$ *von A positiv.*

Beweis. [46]
Seien $A = (a_{ij})$ und $A^{-1} = (\alpha_{ij})$.
(1) Wegen $A = A^H = \bar{A}^T$ sind alle Diagonalelmente a_{ii} reell.
(2) Die Inverse A^{-1} existiert. Wäre das nicht so, hätte das homogene LGS $Ax = 0$ eine nichttriviale Lösung x. Somit wäre auch $x^H Ax = 0,\ x \neq 0$, was der positiven Definitheit von A widerspricht.
(3) Es gilt $(A^{-1})^H = (A^H)^{-1} = A^{-1}$, d. h. A^{-1} ist hermitesch.
Mit $0 \neq y \in \mathbb{C}^n$ folgen $x = A^{-1}y \neq 0$ und $y^H A^{-1} y = x^H A^H A^{-1} Ax = x^H Ax > 0$,
d. h. A^{-1} ist positiv definit.
(4) Für die Hauptuntermatrix $\tilde{A}$ gelten natürlich $\tilde{A} = \tilde{A}^H$ sowie die positive Definitheit gemäß $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_{m-k+1})^T \neq 0,\ x = (x_1, x_2, ..., x_n)^T \neq 0$ mit den Komponenten $\tilde{x}_j,\ j = 1, 2, ..., m-k+1$, ab der Stelle k, sonst Nullen, und $\tilde{x}^H \tilde{A} \tilde{x} = x^H Ax > 0$.
Damit folgt insbesondere $a_{ii} > 0$ für alle $i = 1, 2, ..., n$.
(5) Um $\det(\tilde{A}) > 0$ zu zeigen, genügt der Nachweis von $\det(A) > 0$ für beliebiges n.
Das macht man mit vollständiger Induktion bezüglich der Dimension der Matrix.
Für $n = 1$ folgen aus $\bar{x}^T a_{11} x > 0,\ x = (x_1) \neq 0$, die Bedingungen $a_{11} > 0$ und $\det(A) = a_{11} > 0$.
Sei die Behauptung richtig für die Dimension $n - 1$.
Für die n-dimensionale positiv definite und hermitesche Matrix $A^{-1} = (\alpha_{ij})$ folgt nach (4) die Beziehung $\alpha_{ii} > 0$. Aber nach der Cramerschen Regel zur Bestimmumg der Elemente der inversen Matrix gilt

$$\alpha_{11} = \frac{1}{\det(A)} \det \begin{pmatrix} a_{22} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

$$\det(A) = \frac{1}{\alpha_{11}} \underbrace{\det \begin{pmatrix} a_{22} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{n2} & \cdots & a_{nn} \end{pmatrix}}_{>0 \text{ nach Voraussetzung}} > 0.$$

□

Bemerkung 2.1 Die Aussagen bleiben auch gültig für Untermatrizen in der etwas allgemeineren Form

$$\tilde{A} = \begin{pmatrix} a_{i_1,i_1} & a_{i_1,i_2} & \cdots & a_{i_1,i_k} \\ a_{i_2,i_1} & a_{i_2,i_2} & \cdots & a_{i_2,i_k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{i_k,i_1} & a_{i_k,i_2} & \cdots & a_{i_k,i_k} \end{pmatrix}, \quad (i_1, i_2, ..., i_k) \text{ Teilfolge aus } (1, 2, ..., n). \quad (2.18)$$

Satz 2.6 *Sei $A = A^T > 0$ (spd).*
Dann gelten:
(1) $a_{ii} > 0 \ \ \forall\ i,$
(2) $a_{ij}^2 < a_{ii} a_{jj} \ \ \forall\ i \neq j,$
(3) $\exists\, k \ \ mit \ \ a_{kk} = \max\limits_{i,j} |a_{ij}|,$
(4) $\exists\ A^{-1}.$

Beweis.
Zu (1): Es gilt $x^T A x > 0,\ x \neq 0.$
Sei $x = e_i = (0, ..., 0, 1, 0, ..., 0)^T \neq 0$ der i-te Einheitsvektor.
Dann folgt damit $0 < x^T A x = a_{ii}$.

Zu (2): Sei $x = \xi e_i + e_j,\ i \neq j$, e_i der i-te Einheitsvektor, analog e_j und $\xi \in \mathbb{R}$.
Dann erhalten wir

$$0 < x^T A x = a_{ii}\xi^2 + 2a_{ij}\xi + a_{jj} \ \text{ für alle } \ \xi.$$

Als quadratische Gleichung in der Variablen ξ kann sie also nur komplexe Nullstellen haben. Ihr Radikand ist

$$\left(\frac{a_{ij}}{a_{ii}}\right)^2 - \frac{a_{jj}}{a_{ii}} \ < \ 0, \quad a_{ii} > 0,$$

$$\text{d. h. } \ a_{ij}^2 - a_{ii}a_{ij} \ < \ 0.$$

Zu (3): Angenommen, dass das betragsgrößte Matrixelement nicht auf der Diagonalen liegt, d. h.

$$\max_{i,j} |a_{ij}| = |a_{k_1 k_2}|,\ k_1 \neq k_2.$$

Dann erhält man wegen $|a_{k_1k_2}| \geq a_{k_1k_1}$, $|a_{k_1k_2}| \geq a_{k_2k_2}$ und Teil (2)

$$a_{k_1k_1}a_{k_2k_2} \leq |a_{k_1k_2}|^2 = a_{k_1k_2}^2 < a_{k_1k_1}a_{k_2k_2}$$

einen Widerspruch.
Zu (4): Der Nachweis ist als Teil (2) des Beweises von Satz 2.5 zu finden. □

Hat die Matrix A spezielle Eigenschaften wie
– Diagonaldominanz,
– Symmetrie,
– Definitheit und/oder
– $a_{ii} > 0$,
so stellt sich auch die Frage, ob diese sich auf die reduzierten Matrizen übertragen. Hier können mehrere Aussagen getroffen werden.
Zunächst betrachten wir im Zusammenhang mit der positiven Definitheit $x^T Ax > 0$ die quadratische Form

$$Q(x) \;=\; x^T Ax \;=\; \sum_{i,j=1}^{n} a_{ij} x_i x_j. \tag{2.19}$$

Unter der Voraussetzung $A = A^T$ und $a_{11} > 0$ nehmen wir unter Berücksichtigung der Formeln (2.11) und (2.12) die folgende Umformung vor.

$$\begin{aligned}
Q(x) &= a_{11}x_1^2 + \sum_{i=2}^{n} a_{i1}x_i + \sum_{j=2}^{n} a_{1j}x_j + \sum_{i,j=2}^{n} a_{ij}x_i x_j \\
&= a_{11}x_1^2 + 2\sum_{i=2}^{n} a_{i1}x_i + \sum_{i,j=2}^{n} a_{ij}x_i x_j, \quad a_{i1} = a_{1i} \\
&= a_{11}\Big(x_1^2 + 2\sum_{i=2}^{n} \frac{a_{i1}}{a_{11}}x_i\Big) + \sum_{i,j=2}^{n} a_{ij}x_i x_j \\
&= a_{11}\Big(x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{a_{11}}x_i\Big)^2 - a_{11}\Big(\sum_{i=2}^{n} \frac{a_{i1}}{a_{11}}x_i\Big)^2 + \sum_{i,j=2}^{n} a_{ij}x_i x_j \\
&= a_{11}\Big(x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{a_{11}}x_i\Big)^2 - a_{11}\sum_{i,j=2}^{n} \frac{a_{i1}a_{j1}}{a_{11}^2}x_i x_j + \sum_{i,j=2}^{n} a_{ij}x_i x_j \\
&= a_{11}\Big(x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{a_{11}}x_i\Big)^2 + \sum_{i,j=2}^{n} \Big(a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}\Big)x_i x_j \\
&= a_{11}\Big(x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{a_{11}}x_i\Big)^2 + \sum_{i,j=2}^{n} a_{ij}^{(1)}x_i x_j.
\end{aligned}$$

Damit ist

$$Q(x) \;=\; a_{11}\Big(x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{a_{11}}x_i\Big)^2 + Q^{(1)}(x^{(1)}), \tag{2.20}$$

wobei

$$Q^{(1)}(x^{(1)}) \;=\; \sum_{i,j=2}^{n} a_{ij}^{(1)}x_i x_j \;=\; x^{(1)T}A^{(1)}x^{(1)}, \quad x^{(1)} = (x_2, x_3, ..., x_n)^T.$$

Satz 2.7 *Sei $A = A^T$ und $a_{ii} > 0$.*
A ist positiv definit gdw. die reduzierte Matrix $A^{(k)}$ gemäß (2.11) positiv definit ist.

Beweis. Es genügt der Nachweis für $A^{(1)}$.
1. $\Rightarrow$:
Sei $x^{(1)} = (x_2, x_3, ..., x_n)^T \neq 0$. Mit $a_{11} > 0$ definiert man $x_1 = -\sum_{i=2}^{n} \frac{a_{i1}}{a_{11}} x_i$, womit auch $x = (x_1, x_2, x_3, ..., x_n)^T \neq 0$ ist. Es gilt wegen $A = A^T > 0$ und mit (2.12) die Symmetrie von $A^{(1)}$ (siehe auch Satz 2.9) sowie mit (2.20)

$$0 < x^T A x = Q(x) = a_{11}\Big(x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{a_{11}} x_i\Big)^2 + x^{(1)T} A^{(1)} x^{(1)} = x^{(1)T} A^{(1)} x^{(1)},$$

was die positive Definitheit von $A^{(1)}$ bedeutet.

2. $\Leftarrow$:
Sei $A^{(1)}$ positiv definit und $x = (x_1, x_2, ..., x_n)^T \neq 0$. Wegen

$$Q(x) \;=\; a_{11}\Big(x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{a_{11}} x_i\Big)^2 + x^{(1)T} A^{(1)} x^{(1)} \;\geq\; 0$$

erhalten wir zunächst die positive Semidefinitheit von A.
Die quadratische Form $Q(x)$ kann aber nur Null werden, falls beide Summanden verschwinden. D. h. $x^{(1)T} A^{(1)} x^{(1)} = 0$, was zur Folge hat, dass $x^{(1)} = (x_2, ..., x_n)^T = 0$ ist. Damit vereinfacht sich der erste Summand zu $a_{11} x_1^2$, der bei $a_{11} > 0$ nur verschwinden kann, wenn $x_1 = 0$ ist. Damit ist $x = 0$, was letztendlich $A > 0$ bedeutet. □

Satz 2.8 *Sei A diagonaldominant mit*

$$a_{ii} \;>\; \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, 2, ..., n. \tag{2.21}$$

Dann haben die reduzierten Matrizen $A^{(k)}$ gemäß (2.11) dieselbe Eigenschaft.

Beweis. Es genügt, den Beweis für $A^{(1)}$ zu führen.

$$A^{(1)} = (a_{ij}^{(1)}), \quad a_{ij}^{(1)} \;=\; a_{ij} - \frac{a_{i1} a_{1j}}{a_{11}}, \quad i, j = 2, 3, ..., n.$$

(a) Zunächst zeigen wir, dass $a_{ii}^{(1)} > 0$ ist.
Wegen $a_{11} > |a_{1i}|,\ a_{ii} > |a_{i1}|$ ist $a_{ii} a_{11} - a_{i1} a_{1i} > 0$ und

$$a_{ii}^{(1)} = a_{ii} - \frac{a_{i1} a_{1i}}{a_{11}} > 0.$$

(b) Dann ist nachzuweisen, dass

$$a_{ii}^{(1)} \;>\; \sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}^{(1)}|, \quad i = 2, 3, ..., n,$$

was hier nur für $i = 2$ gemacht wird, da es für die anderen Zeilen analog verläuft. Die zu zeigende Ungleichung ist also

$$a_{22}^{(1)} \;>\; \sum_{j=3}^{n} |a_{2j}^{(1)}|$$

oder ausgeschrieben

$$a_{22} - \frac{a_{21}a_{12}}{a_{11}} \;>\; \left|a_{23} - \frac{a_{21}a_{13}}{a_{11}}\right| + \left|a_{24} - \frac{a_{21}a_{14}}{a_{11}}\right| + ... + \left|a_{2n} - \frac{a_{21}a_{1n}}{a_{11}}\right|.$$

Falls $a_{21} = 0$ ist, stimmt das wegen (2.21) sofort. Sonst macht man eine kleine Rechnung unter Verwendung der Betragseigenschaften $u - |v| \leq u - v,\ |u| + |v| \geq |u - v|$.

$$\begin{aligned}
a_{11} &> |a_{12}| + |a_{13}| + ... + |a_{1n}|,\\
a_{22} &> |a_{21}| + |a_{23}| + ... + |a_{2n}|,\\
1 &> \frac{|a_{12}|}{a_{11}} + \frac{|a_{13}|}{a_{11}} + ... + \frac{|a_{1n}|}{a_{11}},\\
|a_{21}| &> \frac{|a_{21}a_{12}|}{a_{11}} + \frac{|a_{21}a_{13}|}{a_{11}} + ... + \frac{|a_{21}a_{1n}|}{a_{11}},\\
|a_{21}| - \frac{|a_{21}a_{12}|}{a_{11}} &> \frac{|a_{21}a_{13}|}{a_{11}} + ... + \frac{|a_{21}a_{1n}|}{a_{11}},\\
a_{22} - |a_{21}| &> |a_{23}| + ... + |a_{2n}|.
\end{aligned}$$

Die Addition der letzten beiden Ungleichungen ergibt

$$\begin{aligned}
a_{22} - \frac{|a_{21}a_{12}|}{a_{11}} &> |a_{23}| + \frac{|a_{21}a_{13}|}{a_{11}} + ... + |a_{2n}| + \frac{|a_{21}a_{1n}|}{a_{11}}, \text{ folglich}\\
a_{22} - \frac{a_{21}a_{12}}{a_{11}} &> \left|a_{23} - \frac{a_{21}a_{13}}{a_{11}}\right| + ... + \left|a_{2n} - \frac{a_{21}a_{1n}}{a_{11}}\right|,
\end{aligned}$$

was zu zeigen war. □

Die strenge Diagonaldominanz bleibt somit im Reduktionsprozess erhalten, was bei der Lösung eines LGS mittels GA die Diagonalstrategie erlaubt, also keine Pivotstrategie mit ZV erfordert.

Satz 2.9 *Sei $A = A^T$ und für alle reduzierten Matrizen $A^{(k)}$ gemäß (2.11) und (2.12)*

$$a_{kk}^{(k-1)} \neq 0, \ k = 1, 2, ..., n-1.$$

Dann sind auch alle Matrizen $A^{(k)}$ symmetrisch.

Beweis.
Es reicht, den Nachweis unter Verwendung von (2.12) im ersten Schritt zu führen. Mit

$$a_{ij}^{(1)} \ = \ a_{ij}^{(0)} - \frac{a_{i1}^{(0)} a_{1j}^{(0)}}{a_{11}^{(0)}} \ = \ a_{ij} - \frac{a_{i1} a_{1j}}{a_{11}}$$

und $a_{ij} = a_{ji}$ folgt

$$a_{ji}^{(1)} \ = \ a_{ji} - \frac{a_{j1} a_{1i}}{a_{11}} \ = \ a_{ij} - \frac{a_{1j} a_{1i}}{a_{11}} \ = \ a_{ij}^{(1)}.$$

□

Zum Beispiel sind mit der Diagonaldominanz (2.21) die Diagonalelemente $a_{kk}^{(k-1)} \neq 0$.

Satz 2.10 *Sei A diagonaldominant mit der Ungleichung (2.21).*
Dann gelten die folgenden Aussagen.
(1) A ist regulär.
(2) Falls zusätzlich $A = A^T$ gilt, dann ist A positiv definit.
(3) Falls auch A^T streng diagonaldominant ist, dann ist A positiv definit.

Beweis.
Zu (1): Eine Variante des Nachweises führt über den Kreissatz von Gerschgorin 2.28, der die EW $\lambda(A) \neq 0$ und damit die Regularität von A liefert.
Die andere Möglichkeit ist ein indirekter Beweis.
Angenommen, dass $\det(A) = 0$ ist. Dann besitzt das LGS $Ax = 0$ eine nicht triviale Lösung $x = (x_1, x_2, ..., x_n)^T$ mit mindestens einer Nichtnullkomponente.
Nehmen wir die k-te Gleichung des LGS, so erhalten wir

$$\begin{aligned} 0 \ &= \ \sum_{j=1}^{n} a_{kj} x_j, \\ x_k \ &= \ - \sum_{\substack{j=1 \\ j \neq k}}^{n} \frac{a_{kj}}{a_{kk}} x_j \\ &= \ \sum_{j=1}^{n} \tilde{a}_{kj} x_j, \quad \tilde{a}_{kj} = \begin{cases} 0, & \text{falls } j = k, \\ -a_{kj}/a_{kk}, & \text{falls } j \neq k. \end{cases} \end{aligned}$$

Wegen der Diagonaldominanz haben wir für alle $i = 1, 2, ..., n$ die Ungleichungen

$$\begin{aligned} |a_{ii}| &> \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \\ 1 &> \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|, \\ 1 &> \sum_{j=1}^{n} |\tilde{a}_{ij}|. \end{aligned}$$

Sei

$$M = |x_k| = \max_{i=1(1)n} |x_i| > 0.$$

Damit erhält man $M \geq |x_j|$ und die Abschätzungen

$$\begin{aligned} M = |x_k| &= \Big|\sum_{j=1}^{n} \tilde{a}_{kj} x_j\Big| \leq \sum_{j=1}^{n} |\tilde{a}_{kj}|\,|x_j| \\ M \sum_{j=1}^{n} |\tilde{a}_{kj}| &< M \leq \sum_{j=1}^{n} |\tilde{a}_{kj}||x_j|, \\ 0 &< \sum_{j=1}^{n} |\tilde{a}_{kj}|(|x_j| - M) \leq 0 \end{aligned}$$

und somit den Widerspruch. Also ist $\det(A) \neq 0$.

Zu (2): Eine symmetrische Matrix hat nur reelle EW λ_i. Diese sind wegen (2.21) positiv.
Weiter wenden wir den Satz 2.31 an, der die Ähnlichkeitstransformation $Q^T AQ = Q^{-1}AQ = \Lambda = \mathrm{diag}(\lambda_i)$ mit der Orthogonalmatrix Q liefert.
Sei $x \neq 0$. Mit $x = Qy,\ y \neq 0$, folgt

$$x^T Ax = (Qy)^T AQy = y^T Q^T AQy = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2 > 0.$$

Zu (3): Die Beweisidee findet man in [24].
Man betrachtet die quadratische Form (2.19) in ihrer Zerlegung

$$Q(x) = x^T Ax = \sum_{i,j=1}^{n} a_{ij} x_i x_j = \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_i x_j, \quad x \neq 0. \tag{2.22}$$

Voraussetzungen sind

$$a_{ii} > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad a_{ii} > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}|.$$

Nun nehmen wir zwei Abschätzungen von $Q(x)$ nach unten vor, indem wir in der Formel (2.22) den ersten positiven Summanden verkleinern und den zweiten subtrahieren.

$$\begin{aligned} Q(x) &> \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| x_i^2 - \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|x_i||x_j| \\ &> \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|x_i|(|x_i| - |x_j|), \end{aligned}$$

$$\begin{aligned} Q(x) &> \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}| x_i^2 - \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|x_i||x_j| \\ &> \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}| x_i^2 - \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}|\,|x_j||x_i| \\ &> \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}|\,|x_i|(|x_i| - |x_j|) \\ &> \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|x_j|(|x_j| - |x_i|). \end{aligned}$$

In der Summe beider folgt

$$\begin{aligned} 2Q(x) &> \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,(|x_i|^2 - 2|x_i||x_j| + |x_j|^2) \\ &> \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,(|x_i| - |x_j|)^2 \;\geq\; 0, \end{aligned}$$

also $Q(x) > 0$. □

Die Diagonaldominanz (2.21) allein reicht jedoch nicht aus für die positive Definitheit von A.
Dazu konstruieren wir ein einfaches Gegenbeispiel mit einer (2×2)-Matrix. Sei

$$A = \begin{pmatrix} 1 & \alpha \\ 2\alpha & 2 \end{pmatrix}, \quad \alpha \in (-1,1),\ \alpha \neq 0.$$

Es gilt (2.21), aber spaltenweise nicht für $|\alpha| \geq 1/2$.
Die quadratische Form (2.19) ist

$$Q(x) = x_1^2 + 3\alpha x_1 x_2 + 2x_2^2 = \left(x_1 + \frac{3\alpha}{2} x_2\right)^2 - \frac{9\alpha^2 - 8}{4} x_2^2.$$

Für $(x_1, x_2) = (x_1, -\frac{2}{3\alpha} x_1) \neq 0$ und $|\alpha| > \frac{2\sqrt{2}}{3}$ ist $Q(x) < 0$,
z. B. bei $\alpha = \frac{19}{20},\ x = (1, -\frac{40}{57})^T \neq 0$ ist $Q(x) = -\frac{49}{3249}$.

Satz 2.11 *Sei die Matrix $A = (a_{ij})_{i,j=1}^n$ reell und irreduzibel diagonaldominant. Dann gelten folgende Aussagen:*

(1) $a_{ii} \neq 0$ *für alle* $i = 1, 2, ..., n$,

(2) Determinante $\det(A) \neq 0$.

Beweis.
Zu (1): Die Voraussetzung enthält zunächst die Ungleichung

$$|a_{ii}| \ \geq \ \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|,\ i = 1, 2, ..., n,$$

wobei für mindestens einen Index i die Größerbeziehung gilt.
Die Diagonalelemente einer irreduziblen Matrix verschwinden nicht. Wäre ein $a_{ii} = 0$, dann hätte die Matrix A die Nullzeile a_i. Wir bilden mit dem Permutationsvektor $p = (1, 2, ..., i-1, n, i+1, ..., n-1, i)$ die Permutationsmatrix P und können die Transformation

$$\tilde{A} \ = \ PAP^T \ = \ \begin{pmatrix} — & a_1 & — \\ & \vdots & \\ — & a_n & — \\ & \vdots & \\ — & a_i & — \end{pmatrix} \ = \ \left(\begin{array}{c|c} \tilde{A}_{11} & \tilde{A}_{12} \\ \hline 0 & 0 \end{array}\right)$$

durchführen. Der Außendiagonalblock $\tilde{A}_{21}$ enthält nur Nulleinträge und das würde dann die Reduzibilität der Matrix A bedeuten.
Somit gilt $a_{ii} \neq 0,\ i = 1, 2, ..., n$.

Zu (2): Die Begründung erfolgt damit, dass eine irreduzibel diagonaldominante Matrix regulär ist, womit dann $\det(A) \neq 0$ ist.
Was würde aus der Annahme folgen, dass $\det(A) = 0$ ist?
Dann besitzt das homogene LGS $Au = 0,\ A = (a_{ij}),\ a_{ii} \neq 0$ (aus Teil (1)), eine nicht triviale Lösung u und ihre Komponenten genügen der Beziehung

$$u_i \ = \ - \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{a_{ij}}{a_{ii}} u_j, \ i = 1, 2, ..., n.$$

Weiter haben wir mit

$$\tilde{a}_{ij} = \begin{cases} 0, & \text{falls } j = i, \\ -a_{ij}/a_{ii}, & \text{falls } j \neq i, \end{cases}$$

die schwache Diagonaldominanz in der Gestalt

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |\tilde{a}_{ij}| \ \leq \ 1, \ i = 1, 2, ..., n,$$

wobei $<$ für mindestens einen Index i^* gilt.
Für die Lösung $u \neq 0$ definieren wir die Größe

$$M = |u_k| = \max_{i=1(1)n} |u_i| > 0$$

und erhalten damit die Abschätzungen

$$M \sum_{j=1}^{n} |\tilde{a}_{kj}| \ \leq \ M \ = \ \Big| \sum_{j=1}^{n} \tilde{a}_{kj} u_j \Big| \ \leq \ \sum_{j=1}^{n} |\tilde{a}_{kj}||u_j|,$$

$$0 \ \leq \ \sum_{j=1}^{n} |\tilde{a}_{kj}|(|u_j| - M).$$

Um mit $M \geq |u_j|$ dies zu garantieren, müssen alle $|u_j|$ gleich M sein, insbesondere auch $|u_{i^*}| = M$. Für i^* haben wir jedoch

$$M \sum_{j=1}^{n} |\tilde{a}_{i^*j}| \ < \ M \ \leq \ \sum_{j=1}^{n} |\tilde{a}_{i^*j}||u_j|,$$

$$0 \ < \ \sum_{j=1}^{n} |\tilde{a}_{i^*j}|(|u_j| - M) \ \leq \ 0$$

und somit den Widerspruch. □

Für die spd-Eigenschaft der Matrix A braucht man natürlich u. a. die Bedingung $a_{ii} > 0$ für alle $i = 1, 2, ..., n$.

Satz 2.12 *Sei die Matrix $A = (a_{ij})_{i,j=1}^n$ reell, symmetrisch und irreduzibel diagonaldominant mit $a_{ii} > 0 \; \forall \; i$. Dann ist A positiv definit.*

Beweis.
Es gilt

$$a_{ii} \;\geq\; \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \;\; i = 1, 2, ..., n,$$

wobei für mindestens einen Index i die Größerbeziehung vorliegt.
Für beliebiges $x \in \mathbb{R}^n$ haben wir zunächst

$$\sum_{i=1}^{n} a_{ii} x_i^2 \geq \sum_{i=1}^{n} \Big(\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \Big) x_i^2.$$

Nun betrachten wir die quadratische Form (2.19)

$$Q(x) = \sum_{i,j=1}^{n} a_{ij} x_i x_j = \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} x_i x_j$$

und schätzen sie nach unten ab.

$$\begin{aligned}
Q(x) \;&\geq\; \sum_{i=1}^{n} \Big(\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \Big) x_i^2 - \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \, |x_i||x_j| \\
&=\; \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \, |x_i|(|x_i| - |x_j|), \quad \text{und wegen } A = A^T \\
&=\; \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \, |x_j|(|x_j| - |x_i|), \\
2Q(x) \;&=\; Q(x) + Q(x) \\
&\geq\; \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \, (|x_i| - |x_j|)^2 \;=\; q(x) \;\geq\; 0,
\end{aligned}$$

also $Q(x) \geq 0$.
In welchen Fällen kann $Q(x)$ Null werden? Wir zeigen, dass nur $Q(0)$ in Frage kommt.

Da A irreduzibel ist, können nicht alle $a_{ij},\ i \neq j$, gleichzeitig verschwinden.
Im Fall $|x_1| = |x_2| = ... = |x_n| = c \neq 0$ ist

$$Q(x) = c^2 \Big[\sum_{i=1}^{n} a_{ii} + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (\pm a_{ij}) \Big] > 0$$

wegen der einen Teilsumme i^*, wo

$$a_{i^*i^*} > \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i^*j}|,$$

während für die anderen Teilsummen $\geq$ gilt.
Bleibt noch die Situation übrig, wo $|x_k| \neq |x_j|,\ k \neq j$, ist, aber der zugehörige Faktor $a_{kj} = 0$ ist. O.B.d.A. kann $|x_k| \neq 0$ genommen werden.
Wir bilden zwei Indexmengen in $W = \{1, 2, ..., n\}$. Sei $S_1 = \{i\ :\ |x_i| = |x_k| \neq 0\}$ und $S_2 = W - S_1$. Da A irreduzibel ist, gilt nach Definition 2.13, dass $a_{ij} = 0$ für alle $i \in S_1$ und $j \in S_2$ nicht eintreten kann. Also muss mindestens ein Element $a_{ij} \neq 0$ sein. Dieses bildet aber in der Summe $q(x)$ einen positiven Summanden $|a_{ij}|(|x_i| - |x_j|)^2$, so dass dann $Q(x) > 0$ gilt. □

Bemerkung 2.2 (1) Der Beweis ist etwas anders und länger geführt worden.
Kürzer ist: Eine symmetrische Matrix lässt sich mit einer orthogonalen Transformation (Ähnlichkeitstransformation) diagonalisieren (Satz 2.31), hat also reelle EW und wegen

$$a_{ii} \ \geq \ \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|,\ i = 1, 2, ..., n,$$

und dem Kreissatz von Gerschgorin sind die EW von A nichtnegativ. Mit Satz 2.11 haben wir die Regularität von A und $\lambda(A) \neq 0$, also $\lambda(A) > 0$. Wie im Satz 2.10, Teil (2), folgt die positive Definitheit.

(2) Die Umkehrung von Satz 2.12 ist nicht erfüllt.
Die Matrix

$$\begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix}$$

ist irreduzibel und spd, aber es liegt keine Diagonaldominanz vor.

(3) Ein typisches Beispiel für die Anwendung des Satzes ist die Tridiagonalmatrix tridiag$(-1, 2, -1)$.

(4) Die Voraussetzung der Irreduzibilität im Satz 2.12 ist wichtig. Die Matrix

$$\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

ist offensichtlich reduzibel, während alle anderen Eigenschaften vorhanden sind. Die quadratische Form

$$Q(x) = x^T Ax = (x_1 - x_2)^2 + x_3^2 \begin{cases} = 0, & \text{falls } x_3 = 0,\ x_1 = x_2, \\ > 0, & \text{sonst,} \end{cases}$$

ist nicht positiv definit.

Satz 2.13 *Sei die Matrix $A = (a_{ij})_{i,j=1}^n$ reell und symmetrisch mit $a_{ii} > 0\ \forall\ i$. Wenn A die $\binom{n}{2}$ Bedingungen*

$$a_{ii}a_{kk} \ > \ \Big(\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \Big) \Big(\sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}| \Big), \quad 1 \leq i < k \leq n, \tag{2.23}$$

erfüllt sind, dann ist sie positiv definit.

Die Voraussetzungen in diesem Satz sind sehr streng. So werden sie z. B. nicht von den spd Matrizen aus Bemerkung 2.2 (2) und (3) erfüllt.

Den Prozess der Matrixreduktion haben wir in Definition 2.20 beschrieben. Die Berechnung der Matrixelemente nach Formel (2.12) geht auch ein in die rekursive Darstellung der quadratischen Form (2.20).
Es gelingt uns, diese als Summe vollständiger Quadrate zu notieren.

$$\begin{aligned} Q(x) &= a_{11}\Big(x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{a_{11}} x_i\Big)^2 + Q^{(1)}(x^{(1)}) \\ &= \Big(\sqrt{a_{11}}\, x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{\sqrt{a_{11}}} x_i\Big)^2 + \sum_{i,j=2}^{n} a_{ij}^{(1)} x_i x_j \\ &= \Big(\sqrt{a_{11}}\, x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{\sqrt{a_{11}}} x_i\Big)^2 + \Big(\sqrt{a_{22}^{(1)}}\, x_2 + \sum_{i=3}^{n} \frac{a_{i2}^{(1)}}{\sqrt{a_{22}^{(1)}}} x_i\Big)^2 + \\ &\qquad + \sum_{i,j=3}^{n} a_{ij}^{(2)} x_i x_j \\ &= \ ... \\ &= \Big(\sqrt{a_{11}}\, x_1 + \sum_{i=2}^{n} \frac{a_{i1}}{\sqrt{a_{11}}} x_i\Big)^2 + \Big(\sqrt{a_{22}^{(1)}}\, x_2 + \sum_{i=3}^{n} \frac{a_{i2}^{(1)}}{\sqrt{a_{22}^{(1)}}} x_i\Big)^2 + \\ &\qquad + ... + \Big(\sqrt{a_{nn}^{(n-1,n-1)}}\, x_n\Big)^2. \end{aligned} \tag{2.24}$$

Beispiel 2.11 Die Umformung der spd Matrix

$$A = A^{(0)} = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix}$$

nach (2.24) ergibt

$$\begin{aligned} Q(x) &= 3x_1^2 + 4x_1x_2 + 4x_2x_3 + 3x_2^2 + 4x_2x_3 + 3x_3^2 \\ &= \left(\sqrt{3}\,x_1 + \frac{2}{\sqrt{3}}x_2 + \frac{2}{\sqrt{3}}x_3\right)^2 + \left(\sqrt{\frac{5}{3}}\,x_2 + \frac{2}{\sqrt{15}}x_3\right)^2 + \left(\sqrt{\frac{7}{5}}\,x_3\right)^2. \end{aligned}$$

Sie korrespondiert mit den reduzierten Matrizen

$$A^{(0)},\quad A^{(1)} = \begin{pmatrix} 5/3 & 2/3 \\ 2/3 & 5/3 \end{pmatrix},\quad A^{(2)} = \begin{pmatrix} 7/5 \end{pmatrix}.$$

Diese Vorgehensweise lässt sich nicht auf die Matrix

$$B = B^{(0)} = \begin{pmatrix} 3 & 2 & -2 \\ 2 & 3 & 2 \\ -2 & 2 & 3 \end{pmatrix}$$

anwenden, denn wir erhalten durch Reduktion

$$B^{(1)} = \begin{pmatrix} 5/3 & 10/3 \\ 10/3 & 5/3 \end{pmatrix},\quad B^{(2)} = \begin{pmatrix} -5 \end{pmatrix}.$$

Da aus $b_{33}^{(2)} = -5 < 0$ keine reelle Wurzel ziehbar ist, ist B nicht positiv definit.

Wir fassen das Ergebnis in einem Satz zusammen.

Satz 2.14 *Die quadratische Form $Q(x) = x^T Ax$ von n Variablen zur Matrix $A(n,n)$ ist positiv definit gdw. nach ihrer Reduktion auf eine Summe von Quadrattermen alle zu bildenden Wurzeln reell und positiv sind.*

Wir kommen nun zu einem grundlegenden Satz für die Transformation einer Matrix.

Satz 2.15 Satz von Schur
Gegeben sei eine beliebige Matrix $A \in \mathbb{C}^{n,n}$.
Dann existiert eine unitäre Matrix U $(U^H U = I)$, so dass die Matrix $T = U^H AU$ eine obere (rechte) Dreiecksmatrix ist mit den (komplexen) Diagonalelementen t_{ii} als EW von A.

Beweis. [47]
Der Nachweis erfolgt mittels vollständiger Induktion bez. der Dimension n.
Für eine (1×1)-Matrix, die zugleich eine obere Dreiecksmatrix darstellt, ist die Behauptung mit $U = (1)$ gültig.
Die Aussage möge also für die $((n-1) \times (n-1))$-Matrix gelten und $T = T_{n-1}$ sei die obere Dreiecksmatrix.
Betrachten wir nun die Dimension n.
Sei λ_1 ein EW der Matrix A mit zugehörigem EV u_1, der mit $u_1^H u_1 = 1$ normiert ist, d. h. $Au_1 = \lambda_1 u_1$.
Wir ergänzen $u_1 \neq 0$ mit weiteren Vektoren $v_1, v_2, ..., v_{n-1}$ zu einem Orthonormalsystem (ONS) mittels eines geeigneten Verfahrens (z. B. Gram-Schmidt-Orthogonalisierung). Somit erfüllen diese Vektoren die Beziehungen

$$u_1^H v_j = 0,\ j = 1, 2, ..., n-1, \quad v_i^H v_j = \delta_{ij},\ i, j = 1, 2, ..., n-1.$$

Die Vektoren fassen wir als Spaltenvektoren zu der $(n \times n)$-Matrix

$$U_1 \ = \ \begin{pmatrix} | & | & & | \\ u_1 & v_1 & ... & v_{n-1} \\ | & | & & | \end{pmatrix}$$

zusammen, die der Beziehung $U_1^H U_1 = I$ genügt.
Mit dieser Matrix U_1 bilden wir die Matrix

$$A_1 \ = \ U_1^H A U_1 \ = \ \left(\begin{array}{c|c} \lambda_1 u_1^H u_1 & w^T \\ \hline 0 & B \end{array}\right),$$

wobei in w^T die Elemente $u_1^H A v_i$ und in B die Elemente $v_j^H A v_i$ eingehen.
Gemäß Induktionsvoraussetzung gilt für die $((n-1) \times (n-1))$-Matrix B die Transformation auf die obere Dreiecksmatrix $T_{n-1} = P^H B P$ mit $P^H P = I$.
Sei

$$U_2 \ = \ \left(\begin{array}{c|c} 1 & 0^T \\ \hline 0 & P \end{array}\right).$$

Damit erhalten wir

$$U_2^H U_2 \ = \ \left(\begin{array}{c|c} 1 & 0^T \\ \hline 0 & P^H P \end{array}\right) \ = \ I$$

und weiterhin unter Verwendung von U_1, U_2 das Matrixprodukt

$$\begin{aligned} U_2^H U_1^H A U_1 U_2 &= U_2^H A_1 U_2 \\ &= \left(\begin{array}{c|c} 1 & 0^T \\ \hline 0 & P^H \end{array}\right) \left(\begin{array}{c|c} \lambda_1 & w^T \\ \hline 0 & B \end{array}\right) \left(\begin{array}{c|c} 1 & 0^T \\ \hline 0 & P \end{array}\right) \\ &= \left(\begin{array}{c|c} \lambda_1 & w^T P \\ \hline 0 & P^H B P \end{array}\right) = \left(\begin{array}{c|c} \lambda_1 & w^T P \\ \hline 0 & T_{n-1} \end{array}\right) \\ &= T_n, \end{aligned}$$

und damit die gewünschte Transformation auf die obere Dreiecksmatrix T_n mit der unitären Matrix $U = U_1 U_2$. □

Bemerkung 2.3 Die Matrizen A und T sind wegen $T = U^{-1} A U$ ähnlich und haben somit die gleichen (komplexen) EW t_{ii}. Wenn A nur reelle EW hat, sind die Größen t_{ii} natürlich reell.
Besonderheiten treten bei zusätzlichen Voraussetzungen, wie z. B. A reell, normal, unitär, hermitesch und/oder $A = A^T$, bezüglich der Gestalt von U, T und der EW von A auf.

1. Für A normal oder unitär hat die Matrix T eine (komplexe) Diagonalform.

Beweis. Zunächst gelten

$$\begin{aligned} T^H T &= (U^H A U)^H U^H A U = U^H A^H (U U^H) A U = U^H A^H A U, \\ T T^H &= U^H A U (U^H A U)^H = U^H A (U U^H) A^H U = U^H A A^H U, \end{aligned}$$

und wegen $A A^H = A^H A$ ist auch $T T^H = T^H T$, d. h. T ist normal.

Die Matrix T hat die obere Dreiecksgestalt

$$T = \begin{pmatrix} t_{11} & t_{12} & t_{13} & \cdots & t_{1n} \\ 0 & t_{22} & t_{23} & \cdots & t_{2n} \\ 0 & 0 & t_{33} & \cdots & t_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & t_{nn} \end{pmatrix}.$$

Damit sind

$$T^H = \begin{pmatrix} \bar{t}_{11} & 0 & 0 & \cdots & 0 \\ \bar{t}_{12} & \bar{t}_{22} & 0 & \cdots & 0 \\ \bar{t}_{13} & \bar{t}_{23} & \bar{t}_{33} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \bar{t}_{1n} & \bar{t}_{2n} & \bar{t}_{3n} & \cdots & \bar{t}_{nn} \end{pmatrix},$$

$$T^H T = \begin{pmatrix} t_{11}\bar{t}_{11} & t_{12}\bar{t}_{11} & \cdots & t_{1n}\bar{t}_{11} \\ t_{11}\bar{t}_{12} & t_{12}\bar{t}_{12}+t_{22}\bar{t}_{22} & \cdots & t_{1n}\bar{t}_{12}+t_{2n}\bar{t}_{22} \\ \cdots & \cdots & \cdots & \cdots \\ t_{11}\bar{t}_{1n} & t_{12}\bar{t}_{1n}+t_{22}\bar{t}_{2n} & \cdots & \sum_{j=1}^{n} t_{jn}\bar{t}_{jn} \end{pmatrix},$$

$$TT^H = \begin{pmatrix} \sum_{j=1}^{n} t_{1j}\bar{t}_{1j} & \sum_{j=2}^{n} t_{1j}\bar{t}_{2j} & \cdots & t_{1n}\bar{t}_{nn} \\ \sum_{j=2}^{n} t_{2j}\bar{t}_{1j} & \sum_{j=2}^{n} t_{2j}\bar{t}_{2j} & \cdots & t_{2n}\bar{t}_{nn} \\ \cdots & \cdots & \cdots & \cdots \\ t_{nn}\bar{t}_{1n} & t_{nn}\bar{t}_{2n} & \cdots & t_{nn}\bar{t}_{nn} \end{pmatrix}.$$

Aus der Gleichheit der beiden letzten Matrizen folgen

$$|t_{11}|^2 = t_{11}\bar{t}_{11} = \sum_{j=1}^{n} t_{1j}\bar{t}_{1j} = \sum_{j=1}^{n} |t_{1j}|^2$$

und somit $|t_{1j}|^2 = 0$ und $t_{1j} = 0$ für alle $j = 2, 3, ..., n$.
Im nächsten Schritt vergleicht man

$$t_{12}\bar{t}_{12} + t_{22}\bar{t}_{22} = t_{22}\bar{t}_{22} = |t_{22}|^2$$

und

$$\sum_{j=2}^{n} t_{2j}\bar{t}_{2j} = \sum_{j=2}^{n} |t_{2j}|^2$$

und erhält daraus $t_{2j} = 0$ für alle $j = 3, 4, ..., n$, usw., so dass die komplexe diagonale Form von T entsteht. □

2. Für A hermitesch hat die Matrix T eine reelle Diagonalform.

Beweis.

Eine hermitesche Matrix ist auch normal. Damit hat T zunächst eine Diagonalform. Wir zeigen, dass $T = T^H$ ist, woraus $t_{ii} \in \mathbb{R},\ i = 1, 2, ..., n$, folgt.

$$T = U^H AU = ((U^H AU)^H)^H = (U^H A^H (U^H)^H)^H = (U^H AU)^H = T^H.$$

Mit $U^H AU = T = \mathrm{diag}(t_{11}, t_{22}, ..., t_{nn}),\ UU^H = I$, folgt

$$AU = UT = U\,\mathrm{diag}(t_{11}, t_{22}, ..., t_{nn}),$$

$$A \begin{pmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ t_{11}u_1 & t_{22}u_2 & \cdots & t_{nn}u_n \\ | & | & & | \end{pmatrix},$$

d. h. die Elemente t_{ii} sind natürlich die EW der Matrix A und die Spaltenvektoren von U die zugehörigen (komplexen) EV. □

Somit ist eine hermitesche Matrix A mit einer unitären Matrix U und der Ähnlichkeitstransformation $U^H AU = U^{-1}AU$ diagonalisierbar und die Diagonale enthält die reellen EW. Die (komplexen) EV als Spalten u_i von U sind linear unabhängige, mehr noch, orthogonale EV im Sinne von $u_i^H u_j = \delta_{ij}$ (δ_{ij} Kronecker-Symbol).

3. Wenn die Matrix A reell und symmetrisch ist, dann wählt man im Beweis von Satz 2.15 einen reellen EW λ_1 und zugehörigen reellen EV u_1. Mit u_1 bilden die reellen Vektoren $v_1, v_2, ..., v_{n-1}$ zusammen ein Orthogonalsystem im Sinne von

$$u_1^T v_i = 0, \quad v_i^T v_j = \delta_{ij}.$$

Die Matrix U ist dann ebenfalls reell und orthogonal. Die Transformation hat die Form $U^T AU = U^{-1}AU = \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$.
Mit diesen Betrachtungen findet man auf andere Weise das Ergebnis von Satz 2.3.

Satz 2.16 *Sei $A = A^H$. Dann gilt, dass A genau dann positiv definit (resp. positiv semidefinit) ist, wenn die EW $\lambda(A)$ positiv (resp. nicht negativ) sind.*

Beweis. Der Nachweis erfolgt für $A > 0$, da er bei $A \geq 0$ analog verläuft.
1. $\Rightarrow$:
Nach den Sätzen 2.3 und 2.4 sind die EW $\lambda(A)$ reell und ebenso die quadratische Form (x, Ax). Die positive Definitheit von A bedeutet

$$(x, Ax) \;=\; x^H Ax \;>\; 0 \;\text{ für alle } x \;\neq\; 0.$$

Angewendet auf die nicht verschwindenden EV v von A heißt dies

$$0 \;<\; (v, Av) \;=\; \lambda(v, v) \;=\; \lambda v^H v \;=\; \lambda \sum_{i=1}^{n} \bar{v}_i v_i \;=\; \lambda \sum_{i=1}^{n} |v_i|^2$$

und damit $\lambda > 0$.

2. $\Leftarrow$:
Nach Satz 2.15 und Bemerkung 2.3 existiert eine unitäre Matrix U, so dass die Matrix A mit einer Ähnlichkeitstransformation auf die reelle Diagonalform überführt werden kann. Das bedeutet

$$\Lambda \;=\; \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_n) \;=\; U^H AU \;=\; U^{-1}AU.$$

Wegen $\lambda_i > 0$ ist die Matrix Λ positiv definit, und es gilt mit $x \neq 0$ sowie wegen

$$\begin{aligned}
(x, Ax) \;&=\; x^H Ax \\
&=\; (Uy)^H A(Uy), \quad x = Uy,\; y \neq 0 \text{ wegen } \exists\, U^{-1} \\
&=\; y^H U^H AUy \;=\; y^H \Lambda y \;=\; (y, \Lambda y) \\
&=\; \sum_{i=1}^{n} \lambda_i |y_i|^2 \;>\; 0
\end{aligned}$$

auch die positive Definitheit der Matrix A. □

Satz 2.17 *Sei $A = A^H$. Dann folgt aus Matrix A positiv definit auch die positive Definitheit ihres Realteils.*

Beweis.
Sei $A = \Re(A) + \imath\Im(A) = B + \imath C,\ B, C$ reell, $\ v = u + \imath w \neq 0,\ u, w \in \mathbb{R}^n$.
Aus $A = A^H$ folgt $B = B^T$ und $(w, Bu) = (u, Bw)$.
Nach Satz 2.4 ist die quadratische Form (v, Av) reell.
Wir betrachten nun

$$\begin{aligned}(v, Bv) &= (u + \imath w, B(u + \imath w)) \ \text{ und wegen } (x, y) = x^H y \\ &= (u, Bu) + (w, Bw) + \imath\underbrace{[-(w, Bu) + (u, Bw)]}_{=0} \\ &= (u, Bu) + (w, Bw),\end{aligned}$$

d. h. (v, Bv) ist reell.
Bleibt noch zu zeigen, dass für reelles $x \neq 0$ die quadratische Form $(x, Bx) > 0$ ist.
Es gilt

$$(x, Ax) \in \mathbb{R} \text{ und } \ 0 < (x, Ax) = (x, (B + \imath C)x) = (x, Bx) + \imath(x, Cx) = (x, Bx). \quad \square$$

Satz 2.18 *Sei A hermitesch und positiv definit. Dann gibt es eine eindeutige Matrix B, ebenfalls hermitesch und positiv definit, mit $B^2 = A$.*

Beweis.
Zunächst kann man die Matrix A gemäß Satz 2.15 mit einer unitären Matrix U, $U^H = U^{-1}$, diagonalisieren.

$$U^H A U = U^{-1} A U = \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, ..., \lambda_n), \quad \lambda_i > 0.$$

Dann ist $\Lambda^{1/2} = \operatorname{diag}(\sqrt{\lambda}_1, \sqrt{\lambda}_2, ..., \sqrt{\lambda}_n)$.
Die Matrix B kann man definieren als $\ B = U\Lambda^{1/2}U^{-1} = U\Lambda^{1/2}U^H$.
– B ist hermitesch, was leicht zu erkennen ist.
– B ist positiv definit, weil mit $v \neq 0,\ y = U^{-1}v \neq 0$ gilt

$$(v, Bv) = v^H U\Lambda^{1/2}U^{-1}v = (U^H v)^H U\Lambda^{1/2}U^{-1}v = y^H\Lambda^{1/2}y = \sum_{i=1}^{n} \sqrt{\lambda}_i |y_i|^2 > 0.$$

– Es gilt $\ B^2 = U\Lambda^{1/2}U^{-1}U\Lambda^{1/2}U^{-1} = U(\Lambda^{1/2})^2U^{-1} = U\Lambda U^{-1} = A$.
– B ist eindeutig bestimmt.
Wäre C eine andere Matrix mit denselben Eigenschaften, so bleibt wegen

$$C^2 = A = U\Lambda U^H = U\Lambda^{1/2}\Lambda^{1/2}U^H = U\Lambda^{1/2}\underbrace{V^{-1}V}_{=I}\Lambda^{1/2}U^H$$

nur die Lösung $V^{-1} = U^{-1} = U^H$ übrig, d. h. $C = B$. $\square$

Anstelle der positiven Definitheit von A und B kann man in der These auch positiv semidefinite Matrizen ($x^H Ax \geq 0\ \forall\ x$) verwenden.

Wir haben schon den Begriff der Ähnlichkeit von Matrizen verwendet.

Definition 2.24 Ähnlichkeitstransformation

Darunter versteht man die Transformation der Matrix A mittels einer regulären Matrix T (Ähnlichkeitsmatrix) auf die Matrix

$$B = T^{-1}AT. \tag{2.25}$$

Die Matrizen A und B heißen ähnlich.
Falls T orthogonal ist, also $T^TT = TT^T = I$ gilt, dann spricht man von einer orthogonalen Ähnlichkeitstransformation $B = T^TAT$.

Satz 2.19 Eigenschaften der Ähnlichkeitstransformation

(1) Ähnliche Matrizen A und $B = T^{-1}AT$ haben die gleichen EW λ sowie die zugehörigen EV x bzw. $T^{-1}x$.
(2) Besitzt A paarweise verschiedene reelle EW λ_i, so ist A ähnlich zu $\Lambda = \mathrm{diag}(\lambda_i)$.

Beweis.
Zu (1): Aus dem EWP $Ax = \lambda x,\ x \neq 0$, folgt mit einer regulären Matrix T

$$y = T^{-1}x \neq 0,\quad By = T^{-1}ATy = T^{-1}Ax = T^{-1}\lambda x = \lambda y, \tag{2.26}$$

Zu (2): Im Kap. 2.2.2 Punkt 5 wird die lineare Unabhängigkeit aller zugehörigen EV x_i nachgewiesen. Deshalb ist die Modalmatrix $X = (x_1, x_2, ..., x_n)$ regulär, und mit $Ax_i = \lambda_i x_i$ und $AX = X\Lambda$ gilt $X^{-1}AX = \Lambda$. □

Die EV x_i bilden im Allgemeinen kein Orthogonalsystem.

Wenn die reelle Matrix A komplexe EW besitzt, dann kann man sie gemäß Satz 2.15 mit einer unitären Matrix U und der Ähnlichkeitstransformation $U^HAU = U^{-1}AU$ auf eine komplexe obere Dreiecksmatrix bringen (einige ihrer Diagonalelemente sind komplex). Es wird aber nicht möglich sein, mit einer orthogonalen Matrix T eine ähnliche reelle obere Dreiecksmatrix zu finden. Deshalb muss mit T das Ergebnis der Ähnlichkeitstransformation allgemeiner angesetzt werden.

Satz 2.20 *Sei $A(n, n)$ eine beliebige reelle Matrix.*
Dann existiert eine orthogonale Matrix T, so dass man die Quasidreiecksgestalt

$$T^{-1}AT = \begin{pmatrix} R_{11} & R_{12} & \dots & R_{1m} \\ & R_{22} & \dots & R_{2m} \\ & & \ddots & \vdots \\ 0 & & & R_{mm} \end{pmatrix} \tag{2.27}$$

mit Blockmatrizen R_{ij} erhält. Die quadratischen Diagonalblöcke $R_{ii},\ i = 1, 2, ..., m$, sind entweder (1×1)-Matrizen, woraus sich jeweils reelle EW ergeben, oder (2×2)-Matrizen, aus denen Paare konjugiert komplexer EW abgeleitet werden können.

Beweis. [42]

Beispiel 2.12 Für die Matrix

$$A = A^T = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$$

führen wir vier Ähnlichkeitstransformationen $B = T^{-1}AT$ durch.

$$T = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},\ T^{-1} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \ \Rightarrow\ B_1 = T^{-1}AT = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix},$$

$$B_2 = TAT^{-1} = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix},$$

$$Q = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix},\ Q^{-1} = Q^T \ \Rightarrow\ B_3 = Q^TAQ = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

Die drei Transformationsmatrizen enthalten in ihren Spalten jeweils die EV, so dass das Ergebnis in jedem Fall $B_i = \Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ mit den beiden EW 2 und 4 ist. Da sie verschieden sind, sind die EV orthogonal. Die vierte Variante ist

$$T = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},\ T^{-1} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \ \Rightarrow\ B_4 = T^{-1}AT = \begin{pmatrix} 2 & -1 \\ 0 & 4 \end{pmatrix}.$$

Die Matrix T enthält in der ersten Spalte den EV $(1,1)^T$ zum EW 2, aber ihre zweite Spalte ist kein EV. Glücklicherweise hat die Ähnlichkeitstransformation zu einer oberen Dreiecksmatrix B_4 geführt, so dass die EW von A dann als Diagonalelemente sichtbar sind.

Beispiel 2.13 Wir nehmen die Matrix

$$A = \begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix} = \begin{pmatrix} \frac{39}{50} & \frac{563}{1000} \\ \frac{913}{1000} & \frac{659}{1000} \end{pmatrix}$$

aus Beispiel 2.10 mit der Betrachtung ihres EWP.
Die EW sind

$$\lambda_{1,2} = \frac{1439}{2000} \pm \frac{\sqrt{2070717}}{2000} = 1.4389993050726317415,\ 0.69492736825854 \cdot 10^{-6} > 0,$$

die zugehörigen linear unabhängigen skalierten EV

$$\begin{aligned} \tilde{v}_1 &= \Big(1, -\frac{121}{1126} + \frac{\sqrt{2070717}}{1126}\Big)^T = (1,\ 1.1705138633616904822)^T, \\ \tilde{v}_2 &= \Big(1, -\frac{121}{1126} - \frac{\sqrt{2070717}}{1126}\Big)^T = (1, -1.3854339344096478534)^T. \end{aligned}$$

Nach Normierung mittels euklidischer Norm sind die EV

$$v_1 = c_1\Big(1, -\frac{121}{1126} + \frac{\sqrt{2070717}}{1126}\Big)^T = \begin{pmatrix} 0.64955572831439685710 \\ 0.76031398501800126831 \end{pmatrix},$$

$$c_1 = \Big[1 + \Big(-\frac{121}{1126} + \frac{\sqrt{2070717}}{1126}\Big)^2\Big]^{-1/2},$$

$$v_2 = c_2\Big(1, -\frac{121}{1126} - \frac{\sqrt{2070717}}{1126}\Big)^T = \begin{pmatrix} 0.58526314402966098506 \\ -0.81084342029797362184 \end{pmatrix},$$

$$c_2 = \Big[1 + \Big(-\frac{121}{1126} - \frac{\sqrt{2070717}}{1126}\Big)^2\Big]^{-1/2}.$$

Wir machen nun die Orthogonaltransformation von A auf eine obere Dreiecksmatrix R gemäß

$$Q^T A Q = R, \;\; Q = (q_1, q_2), \; Q^T Q = I, \; R = \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix}.$$

Wegen der Ähnlichkeit von A und R ist $r_{11} = \lambda_1$ und $r_{22} = \lambda_2$.
Wegen $(q_1, q_2)R = A(q_1, q_2)$ kann man die erste Spalte von Q zu $q_1 = v_1$ wählen.
Als Vektor $q_2 \perp q_1$ nimmt man einfach $q_2 = (q_{1,2}, -q_{1,1})^T, \; \|q_2\|_2 = 1$.
Aus $r_{12}q_1 + r_{22}q_2 = Aq_2$ und $q_1^T q_2 = 0$ folgt zunächst $r_{12} = q_1^T A q_2 = 0.35$, gleichzeitig aber auch $r_{22} = q_2^T A q_2 = q_2^T A^T q_2$, was zur Kontrolle der Übereinstimmung mit λ_2 dienen kann.
Somit erhält man

$$Q = \begin{pmatrix} 0.64955572831439685710 & 0.76031398501800126831 \\ 0.76031398501800126831 & -0.64955572831439685710 \end{pmatrix},$$

$$R = \begin{pmatrix} 1.4389993050726317415 & 0.35 \\ 0 & 0.69492736825854 \cdot 10^{-6} \end{pmatrix}.$$

Wir notieren noch einige Eigenschaften von reellen Matrizen mit einer besonderen Vorzeichensituation bezüglich ihrer Elemente.

Satz 2.21 *Ist die Matrix $A(n,n) = (a_{ij})$ irreduzibel und sind alle $a_{ij} \geq 0$, dann gelten die folgenden Aussagen.*
(1) Der Spektralradius ist $\rho(A) > 0$ und wächst, wenn irgendein Matrixelement wächst.
(2) Es gibt einen einfachen EW $\lambda(A) = \rho(A)$.
(3) Die Komponenten des dazugehörigen EV $x = (x_1, x_2, ..., x_n)^T$ erfüllen $x_i \neq 0 \; \forall \; i$ und haben einheitliches Vorzeichen.

Beweis. [50]

Satz 2.22 *(1) Eine symmetrische und schwach diagonaldominante* L*-Matrix ist positiv semidefinit.*
(2) Eine symmetrische und irreduzibel diagonaldominante L*-Matrix ist positiv definit.*

Beweis. Zu (1): Es ist

$$A = A^T = A^H > 0, \quad a_{ii} > 0, \quad a_{ii} \geq \sum_{j\,=\,1,\,j\,\neq\,i}^{n} |a_{ij}|, \quad i = 1, 2, ..., n.$$

Nach den Sätzen 2.3, 2.4 und 2.28 haben wir zunächst

$$\text{EW } \lambda(A) \in \mathbb{R},\ (x, Ax) \in \mathbb{R},\ \lambda(A) \geq 0.$$

Weiter gelten nach Satz 2.31 die Transformationsbeziehungen

$$U^T A U = \Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n),\ A = U \Lambda U^T,\ U^T U = I.$$

Somit folgt für die quadratische Form mit $x \in \mathbb{R}^n$

$$\begin{aligned}(x, Ax) \ = \ x^T A x \ &= \ x^T U \Lambda U^T x \\ &= \ (U^T x)^T \Lambda (U^T x),\ y = U^T x \\ &= \ y^T \Lambda y \\ &= \ \sum_{i=1}^{n} \lambda_i y_i^2 \geq 0.\end{aligned}$$

Zu (2): Zunächst gehen wir wie in Teil (1) vor und erhalten $A \geq 0$.
Gemäß Satz 2.11 ist A regulär und $\lambda = 0$ kein EW von A. Somit ist für $x \neq 0$ auch $y \neq 0$ und $(x, Ax) > 0$. □

Eine symmetrische und irreduzibel diagonaldominante L-Matrix ist damit eine Stieltjes-Matrix.

Satz 2.23 *Sei $A = (a_{ij})$ eine* M*-Matrix. Dann gilt:*
(1) A ist eine L*-Matrix,*
(2) die Diagonalelemente a'_{ii} der inversen Matris $A^{-1} = (a'_{ij})$ sind positiv.

Beweis. Zu (1): Es ist nur $a_{ii} > 0$ zu zeigen.
Wegen $a_{ij} \leq 0,\ i \neq j$, $a'_{ij} \geq 0$ und $AA^{-1} = I$, d. h.

$$\begin{aligned}\sum_{j=0}^{n} a_{ij} a'_{ji} \ &= \ 1, \quad i = 1, 2, ..., n, \\ a_{ii} a'_{ii} \ &= \ 1 - \sum_{\substack{j\,=\,1 \\ j\,\neq\,i}}^{n} a_{ij} a'_{ji} \ \geq \ 1\end{aligned}$$

folgt $a_{ii} > 0$ für alle i.
Zu (2): Aus $a_{ii} a'_{ii} \geq 1$ und $a_{ii} > 0$ ergibt sich auch $a'_{ii} > 0$. □

Manchmal findet man die Bedingung $a_{ii} > 0$ auch mit in der Definition der M-Matrix.

Satz 2.24 *Sei A eine* L*-Matrix, $A = D - C$ ihre Zerlegung mit $D = \mathrm{diag}(A) = \mathrm{diag}(d_{11}, d_{22}, ..., d_{nn})$, $C = (c_{ij})$ und $J = D^{-1}C$.*
Dann ist A eine M*-Matrix gdw. der Spektralradius $\rho(J) < 1$ ist.*

Beweis. [25]
1. $\Leftarrow$:
Sei $\rho(J) < 1$. Dann hat die Matrix $I - J$ keinen Null-EW und ist regulär.
Weiterhin konvergiert die unendliche Reihe $I + J + J^2 + ...$, und zwar gegen $(I - J)^{-1}$, d. h.

$$\tilde{J} = (I - J)^{-1} = \sum_{k=0}^{\infty} J^k.$$

Wegen $d_{ii} > 0$ und $c_{ij} \geq 0$ sind alle Matrixelemente von $J = D^{-1}C$ nicht negativ, ebenso die von ihren Potenzen J^k und $\tilde{J}$.
Die Matrix $A = D(I - J)$ ist regulär und ihre Inverse berechnet sich aus

$$A^{-1} = (I - J)^{-1}D^{-1} = \tilde{J}D^{-1} = (a'_{ij})$$

und hat damit nur nicht negative Einträge a'_{ij}.

2. $\Rightarrow$:
Die M-Matrix $A = D - C$ ist gemäß Satz 2.23 auch eine L-Matrix.
Dann kann man leicht überprüfen, dass die Matrix

$$\hat{A} = D^{-1/2}AD^{-1/2} = D^{-1/2}(D - C)D^{-1/2} = I - D^{-1/2}CD^{-1/2},$$

wobei $D^{-1/2} = \mathrm{diag}(d_{11}^{-1/2}, d_{22}^{-1/2}, ..., d_{nn}^{-1/2})$ ist, die Bedingungen (1.1)–(1.2) einer M-Matrix in der Definition 2.16 erfüllt.
Weiter gilt mit $J = D^{-1}C = (J_{ij}),\ J_{ij} \geq 0$,

$$\begin{aligned} \hat{J} &= D^{1/2}JD^{-1/2} = (\hat{J}_{ij}),\quad \hat{J}_{ij} \geq 0, \\ &\quad J \text{ und } \hat{J} \text{ sind ähnlich mit gleichem Spektrum,} \\ &= D^{-1/2}CD^{-1/2}. \end{aligned}$$

Somit sind $\hat{A} = I - \hat{J} = (\hat{a}_{ij})$ und $\hat{A}^{-1} = (I - \hat{J})^{-1} = (\hat{a}'_{ij}),\ \hat{a}'_{ij} \geq 0$.
Wegen der Identität

$$(I - \hat{J})^{-1} = (I + \hat{J} + ... + \hat{J}^m) + (I - \hat{J})^{-1}\hat{J}^{m+1}$$

für alle ganzzahligen Werte $m \geq 0$ untersuchen wir das Verhalten der Matrixterme auf der rechten Seite.

Die Elemente der Matrix $K^{(m)} = I + \hat{J} + ... + \hat{J}^m = (K_{ij}^{(m)})$ bilden mit zunehmendem m schwach monoton wachsende Folgen. Da auch die Elemente des zweiten Summanden nicht negativ sind, gilt

$$0 \leq K_{ij}^{(0)} \leq ... \leq K_{ij}^{(m)} \leq K_{ij}^{(m+1)} \leq \hat{a}'_{ij} \quad \forall\ m.$$

Daher müssen die Folgen $\{K_{ij}^{(m)}\}_{m=0}^{\infty}$ für alle i, j konvergieren, was

$$\lim_{m\to\infty} (K^{(m)} - K^{(m-1)}) \ = \ \lim_{m\to\infty} \hat{J}^m \ = \ 0$$

nach sich zieht.
Aber es ist $\hat{J}^m \to 0$ gdw. für den Spektralradius $\rho(\hat{J}) = \rho(J) < 1$ erfüllt ist. □

Die Matrix $J = D^{-1}C$ tritt später bei der Untersuchung von IV auf und ihre Eigenschaften sind für Konvergenz des IV wichtig.

Satz 2.25
Sei A eine Stieltjes-Matrix, dann existiert $A^{-1} = (a'_{ij})$ mit $a'_{ij} \geq 0$ und $a'_{ii} > 0$.

Beweis. Gemäß Satz 2.6 folgt aus der positiven Definitheit von A ihre Regularität. In Satz 2.5 wurde bewiesen, dass auch $A^{-1} = (a'_{ij})$ positiv definit ist und damit gilt $a'_{ii} > 0$. Weiterhin sind $\det(A) > 0$ und $D = \mathrm{diag}(A) > 0$.
Sei $A = D - C = D(I - D^{-1}C) = D(I - J),\ J = D^{-1}C$. Somit hat man $A^{-1} = (I - J)^{-1}D^{-1}$. Gemäß Satz 2.24 brauchen wir nur noch die Ungleichung $\rho(J) < 1$ zu zeigen.
Nehmen wir an, dass $\rho(J) \geq 1$ ist. Nach F.G. Frobenius [1908] oder Satz 2.21 ist die Größe $\mu = \rho(J)$ EW von J. Weiter haben wir die spd-Eigenschaft der Matrix $\hat{A} = D^{-1/2}AD^{-1/2} = I - \hat{J}$ mit der Ähnlichkeit der Matrizen $\hat{J}$ und J gemäß

$$\hat{J} = D^{-1/2}CD^{-1/2} = D^{1/2}JD^{-1/2}.$$

Somit sind die EW von $\hat{A}$ positiv. Andererseits hat $\hat{A}$ den EW $1 - \mu \leq 0$ und wäre damit nicht positiv definit, was zu einen Widerspruch führt. □

Eine Stieltjes-Matrix ist also eine M-Matrix.

Satz 2.26
Die Matrix A ist monoton gdw. mit einem beliebigen Vektor $x \in \mathbb{R}^n$ aus den Ungleichungen $(Ax)_i \geq 0,\ i = 1, 2, ..., n,$ die Beziehungen $x_i \geq 0$ folgen.

Beweis.
1. $\Rightarrow$:
Aus $y = Ax \geq 0$ (elementweise) folgen mit $A^{-1} = (a'_{ij}),\ a'_{ij} \geq 0$, die Beziehungen $x = A^{-1}y$ und $x_i = \sum_{j=1}^{n} a'_{ij}y_j \geq 0,\ i = 1, 2, ..., n.$

2. $\Leftarrow$:
Zuerst zeigen wir die Regularität von A, indem wir als einzige Lösung des homogenen LGS $Ax = 0$ die Nulllösung finden.
Die Gleichung $Ax = 0$ bedeutet auch $Ax \geq 0$ und somit $x \geq 0$ komponentenweise. Aber $Ax = 0$ heißt ebenfalls $-Ax = A(-x) = 0 \geq 0$ und $-x \geq 0$ komponentenweise. Somit kann nur $x = 0$ sein.
Sei

$$A^{-1} = (a'_{ij}) = \begin{pmatrix} | & | & & | \\ a'_1 & a'_2 & \cdots & a'_n \\ | & | & & | \end{pmatrix}.$$

$AA^{-1} = I$ ist ausgeschrieben

$$\begin{pmatrix} | & | & & | \\ Aa'_1 & Aa'_2 & \cdots & Aa'_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ e_1 & e_2 & \cdots & e_n \\ | & | & & | \end{pmatrix},$$

woraus mit $Aa'_j = e_j,\ j = 1, 2, ..., n$, und elementweise ≥ 0 die Beziehung $a'_j \geq 0$, d. h. $a'_{ij} \geq 0\ \forall\ i, j$ folgt. □

Dieser Satz kann auch als weitere Möglichkeit der Definition einer monotonen Matrix interpretiert werden.
Ganz einfach lassen sich die Aussagen des folgenden Satzes zeigen.

Satz 2.27 *Sei $A = (a_{ij})$.*
(1) Falls A eine M*-Matrix ist, dann ist A monoton.*
(2) Falls A monoton und $a_{ij} \leq 0,\ i \neq j$, sind, dann ist A eine M*-Matrix.*
(3) Falls A monoton und L*-Matrix ist, dann ist A eine* M*-Matrix.*

Beispiel 2.14 Die Matrix

$$A \ = \ A(n,n) \ = \ \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}$$

ist eine typische Matrix, die zahlreiche der genannten Merkmale besitzt.
Das sind solche wie

– schwach besetzt,
– tridiagonal und damit Bandstruktur,
– schwach diagonaldominant,

– irreduzibel diagonaldominant,
– symmetrisch und positiv definit, positive relle EW,
– diagonalisierbar,
– Eigenschaft A,
– konsistent geordnet (Eigenschaft wird später erläutert),
– L-Matrix, Stieltjes-Matrix,
– M-Matrix, monoton, dazu notieren wir die symmetrische inverse Matrix (nur oberes Dreieck)

$$A^{-1} = (a'_{ij}) = \begin{pmatrix} n\frac{1}{n+1} & (n-1)\frac{1}{n+1} & (n-2)\frac{1}{n+1} & \cdots & 2\frac{1}{n+1} & \frac{1}{n+1} \\ & (n-1)\frac{2}{n+1} & (n-2)\frac{2}{n+1} & \cdots & 2\frac{2}{n+1} & \frac{2}{n+1} \\ & & (n-2)\frac{3}{n+1} & \cdots & 2\frac{3}{n+1} & \frac{3}{n+1} \\ & & & \ddots & \vdots & \vdots \\ & & & & 2\frac{n-1}{n+1} & \frac{n-1}{n+1} \\ & & & & & \frac{n}{n+1} \end{pmatrix}$$

mit ihren nicht negativen Elementen $a'_{ij} = i(n+1-j)/(n+1),\ i \le j$, und $a'_{ii} > 0$. Hier sind sogar alle Einträge positiv.

– Mit der regulären Zerlegung $A = D - C,\ D = \mathrm{diag}(A)$, erhält man die Matrix

$$J \;=\; D^{-1}C \;=\; \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \;=\; \mathrm{tridiag}(d, 0, d),\ d = \frac{1}{2}.$$

Ihr Spektralradius ist $\rho(J) < 1$.
Wir notieren auch die EW λ_i und EV $v^{(i)}$ von J.

$$\begin{aligned} \lambda_i &= \cos(i\pi/(n+1)), \quad i = 1, 2, ..., n, \\ &\quad -1 < \lambda_n < \lambda_{n-1} < .... < \lambda_2 < \lambda_1 = \cos(\pi/(n+1)) < 1, \\ \lambda_{n+1-i} &= -\lambda_i, \quad i = 1, 2, ..., [\tfrac{n}{2}], \quad \text{falls } n \text{ ungerade, dann } \lambda_{\frac{n+1}{2}} = 0, \\ v^{(i)} &= (v_1^{(i)}, v_2^{(i)}, ..., v_n^{(i)})^T \text{ mit } v_j^{(i)} = \sin(ij\pi/(n+1)). \end{aligned}$$

2.2 Eigenwertproblem

Wir untersuchen das (lineare) EWP für eine reelle Matrix $A(n,n)$.

2.2.1 Eigenschaften der Eigenwerte

Diese ergeben sich auch im Zusammenhang mit dem charakteristischen Polynom.

1. Die Größe λ ist EW von A genau dann, wenn λ eine Nullstelle des charakteristischen Polynoms $p_n(\lambda)$ ist. Das Polynom hat genau n (komplexe) Nullstellen, zählt man die Vielfachheit der Nullstellen mit.

2. Die Menge der n EW von A heißt **Spektrum** $\sigma(A)$.

$$\sigma(A) \;=\; \{\lambda\} \;=\; \{\lambda_1, \lambda_2, ..., \lambda_n\}. \tag{2.28}$$

Der **Spektralradius** $\rho(A)$ ist

$$\rho(A) \;=\; \max_{\lambda\in\sigma(A)} |\lambda| \;=\; \max_{i=1,2,...,n} |\lambda_i|. \tag{2.29}$$

3. Ein EW λ_i besitzt die **algebraische Vielfachheit** $n_i \geq 1$, falls λ_i eine n_i-fache Nullstelle von $p_n(\lambda)$ ist. Es gilt $\sum_i n_i = n$. λ_i besitzt die **geometrische Vielfachheit** $m_i \geq 1$, falls $m_i = n - \text{rang}(A - \lambda_i I)$ gilt. Man spricht vom Defekt von $A - \lambda_i I$. Es gilt stets $m_i \leq n_i$, wobei m_i die Anzahl der linear unabhängigen EV zu λ_i ist.

4. Mit dem charakteristischen Polynom $p_n(\lambda)$ folgt unter Einbeziehung der Entwicklungsätze für Determinanten

$$\begin{aligned} q_n(\lambda) \;=\; (-1)^n p_n(\lambda) \;&=\; (-1)^n \;\det(A - \lambda I) \;=\; \det(\lambda I - A) \\ &=\; \lambda^n - (a_{11} + a_{22} + ... + a_{nn})\lambda^{n-1} + ... \\ &=\; \lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + ... + c_n \\ &=\; (\lambda - \lambda_1)(\lambda - \lambda_2) \cdot ... \cdot (\lambda - \lambda_n) \\ &=\; \lambda^n - \lambda^{n-1} \sum_{i=1}^{n} \lambda_i + ... + (-1)^n \prod_{i=1}^{n} \lambda_i, \end{aligned} \tag{2.30}$$

und es gelten

(1) $\text{spur}(A) \;=\; \sum_{i=1}^{n} \lambda_i \;=\; -c_1$, wobei $\text{spur}(A) = \sum_{i=1}^{n} a_{ii}$ die Spur des Matrix ist; für die Spur findet man auch die Bezeichnung $\text{trace}(A)$,

(2) $\det(A) \;=\; \prod_{i=1}^{n} \lambda_i \;=\; (-1)^n c_n$,

(3) $\sum_{i=1}^{n} |\lambda_i|^2 \;\leq\; \text{spur}(AA^T) \;=\; \text{spur}(A^TA) \;=\; \|A\|_F^2$ (Frobenius-Norm).

Teil (1) ist leicht nachzuprüfen, die erste Gleichheit in (2) folgt aus 2.30 mit $\lambda = 0$ oder basiert auf Satz 2.15 mit $A = UTU^{-1}$, Teil (3) wird noch als eigenständiger Satz formuliert.

5. Eine singuläre Matrix A, d. h. $\det(A) = \det(A - 0 \cdot I)$, hat also mindestens einen Null-EW. Die dazu linear unabhängigen EV spannen den Nullraum von A auf.

6. Für die Lokalisierung der EW gilt der folgende Satz.

Satz 2.28 Kreissatz von Gerschgorin

Für jeden Eigenwert $\lambda(A)$ von $A = (a_{ij})$ gibt es ein $k \in \{1, 2, ..., n\}$, so dass gilt

$$|\lambda - a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}|. \tag{2.31}$$

Beweis.

Seien $Ax = \lambda x,\ x \neq 0,\ |x_k| = \max\limits_{j=1,...,n} |x_j| > 0,\ B = A - \lambda I = (b_{ij})$.

Wir betrachten die k-te Zeile des LGS $Bx = 0$ und erhalten

$$\begin{aligned} \sum_{j=1}^{n} b_{kj} x_j &= 0, \\ b_{kk} + \sum_{\substack{j=1 \\ j \neq k}}^{n} b_{kj} \frac{x_j}{x_k} &= 0, \\ |b_{kk}| &= \Big| \sum_{\substack{j=1 \\ j \neq k}}^{n} b_{kj} \frac{x_j}{x_k} \Big| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |b_{kj}|, \\ |a_{kk} - \lambda| &\leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}|. \end{aligned}$$

□

Die EW liegen in einem Gebiet (der komplexen Ebene) als Vereinigung von Kreisen jeweils mit den Mittelpunkten a_{ii} und den Radien $r_i = \sum\limits_{j=1,\, j \neq i}^{n} |a_{ij}|$.

Für reelle EW ergeben sich entsprechende Intervalle.

Folgerung 2.29

(1) Da A und A^T wegen $\det(A^T - \lambda I) = \det((A - \lambda I)^T) = \det(A - \lambda I)$ die gleichen EW haben, gilt der Kreissatz auch bezüglich der Spalten.

(2) Ist A streng diagonaldominant und $a_{ii} > 0$, dann kann gemäß Kreissatz der Koordinatenursprung nicht zur Vereinigung der Kreise gehören. Null ist also kein EW und die Matrix ist regulär.

(3) Aus $|\lambda| - |a_{ii}| \leq |\lambda - a_{ii}| \leq \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$ folgt unmittelbar

$$|\lambda| \leq \rho(A) \leq \max_{i=1,...,n} \sum_{j=1}^{n} |a_{ij}| = \|A\|_\infty \quad (Zeilensummennorm) \tag{2.32}$$

(4) Analog gilt $|a_{ii}| - |\lambda| \ \leq \ |\lambda - a_{ii}| \ \leq \ \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|,$

und wegen λ^{-1} *als EW von* A^{-1} *folgt*

$$\frac{1}{\rho(A^{-1})} \ \geq \ \min_{i=1,\ldots,n} \Big(|a_{ii}| - \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \Big). \tag{2.33}$$

(5) Sei $A = A^T > 0$. *Dann gelten die Beziehungen*

$$\begin{aligned} \rho(A) &= \max_{x \neq 0} \frac{x^T A x}{x^T x} \quad \Big(R(x) = \frac{x^T A x}{x^T x} \ \textit{Rayleigh} - \textit{Quotient} \Big), \\ \frac{1}{\rho(A^{-1})} &= \min_{x \neq 0} \frac{x^T A x}{x^T x}. \end{aligned} \tag{2.34}$$

Beweis. Nach Satz 2.15, Bemerkung 2.3 und Satz 2.16 haben wir die n reellen EW $\lambda_n \geq \lambda_{n-1} \geq ... \geq \lambda_1 > 0$ und die zugehörigen reellen orthogonalen EV v_i, die eine Basis im Raum $\mathbb{R}^n$ darstellen.
Für einen beliebigen nicht verschwindenden Vektor $x \in \mathbb{R}^n$ gilt

$$\begin{aligned} x &= \sum_{i=1}^{n} c_i v_i \neq 0, \\ x^T x &= \sum_{i,j=1}^{n} c_i c_j v_i^T v_j \ = \ \sum_{i=1}^{n} c_i^2, \quad v_i^T v_j = \delta_{ij}, \\ x^T A x &= \sum_{i,j=1}^{n} c_i c_j v_i^T A v_j \ = \ \sum_{i=1}^{n} c_i^2 \lambda_i, \quad A v_j = \lambda_j v_j, \\ \frac{x^T A x}{x^T x} &= \frac{\sum\limits_{i=1}^{n} c_i^2 \lambda_i}{\sum\limits_{i=1}^{n} c_i^2} \ = \ \sum_{i=1}^{n} \frac{c_i^2}{\sum\limits_{j=1}^{n} c_j^2} \lambda_i. \end{aligned}$$

Der Ausdruck

$$\frac{x^T A x}{x^T x} \ = \ \sum_{i=1}^{n} w_i \lambda_i, \quad w_i = \frac{c_i^2}{\sum\limits_{j=1}^{n} c_j^2} \geq 0, \quad \sum_{i=1}^{n} w_i = 1,$$

ist eine gewichtete Summe. Für diese gilt

$$\lambda_n \ \geq \ \sum_{i=1}^{n} w_i \lambda_i \ \geq \ \lambda_1.$$

Also sind

$$\lambda_n \;=\; \rho(A) \;\geq\; \max_{x\neq 0}\frac{x^TAx}{x^Tx} \quad \text{und} \quad \lambda_1 \;\leq\; \min_{x\neq 0}\frac{x^TAx}{x^Tx}$$

sowie wegen der EW $\lambda_1^{-1} \geq \lambda_2^{-1} \geq ... \geq \lambda_n^{-1} > 0$ von A^{-1} und $\rho(A^{-1}) = \lambda_1^{-1}$ auch

$$\lambda_1 \;=\; \frac{1}{\lambda_1^{-1}} \;=\; \frac{1}{\rho(A^{-1})} \;\leq\; \min_{x\neq 0}\frac{x^TAx}{x^Tx}.$$

Zusätzlich gelten

$$\lambda_n \;=\; \frac{v_n^TAv_n}{v_n^Tv_n}, \quad \lambda_1 \;=\; \frac{v_1^TAv_1}{v_1^Tv_1}.$$

□

(6) Eine Verallgemeinerung von (5) erhalten wir für $A = A^H$ mit $\lambda(A) \in \mathbb{R}$. Dann gelten die Beziehungen

$$\begin{aligned} &\textit{größter EW von } A && \lambda_{max} \;&=\; \max_{x\neq 0}\frac{x^HAx}{x^Hx}, \\ &\textit{kleinster EW von } A && \lambda_{min} \;&=\; \min_{x\neq 0}\frac{x^HAx}{x^Hx}. \end{aligned} \tag{2.35}$$

Der Nachweis verläuft analog zu (5).

(7) Ist A irreduzibel diagonaldominant und $a_{ii} > 0\;\forall\, i$, dann gilt für die EW $\lambda(A) \neq 0$ und ihr Realteil $\Re(\lambda(A)) > 0$.

Beweis. Nach Satz 2.11 ist $\det(A) = \det(A - 0 \cdot I) \neq 0$ und somit Null kein EW. Gemäß Kreissatz von Gerschgorin ist $\Re(\lambda(A)) \geq 0$. Da Null als EW entfällt, bleibt $\Re(\lambda(A)) > 0$. □

Beispiel 2.15 Anwendung des Kreissatzes von Gerschgorin

$$\text{Gerschgorin-Kreise } \; \mathcal{K}_i = \Big\{\; z \;:\; |z - a_{ii}| \leq r_i, \quad r_i = \sum_{j\,=\,1,\,j\,\neq\,i}^{n} |a_{ij}| \Big\},\; i = 1, 2, ..., n.$$

$$A_1(3,3) = \begin{pmatrix} 12 & -2 & 3 \\ -1 & 8 & -2 \\ -1 & 3 & 12 \end{pmatrix},$$

$$\sigma(A_1) = \{8.136\,836\,124\,478,\; 11.931\,581\,937\,761 \pm 2.542\,970\,427\,631\,\imath\},$$

$$\begin{aligned} \lambda(A_1) \in \bigcup_{i=1}^{3}\mathcal{K}_i &= \{z : |z-12| \leq 5\} \cup \{z : |z-8| \leq 3\} \cup \{z : |z-12| \leq 4\} \\ &= \{z : |z-12| \leq 5\} \cup \{z : |z-8| \leq 3\}. \end{aligned}$$

$$A_2(3,3) = \begin{pmatrix} 5 & -2 & -4 \\ -2 & 2 & 2 \\ -4 & 2 & 5 \end{pmatrix}, \quad A_2 = A_2^T,$$

$\sigma(A_2) = \{1, 1, 10\}$,

$$\lambda(A_2) \in \bigcup_{i=1}^{3} \mathcal{K}_i = \{z : |z-5| \le 6\} \cup \{z : |z-2| \le 4\} \cup \{z : |z-5| \le 6\} = [-2, 11].$$

$$A_3(n,n) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} = \text{tridiag}(1, 0, 1), \quad A_3 = A_3^T,$$

EW: $\lambda_i = 2\cos(ih) \in (-2, 2),\ i = 1, 2, ..., n, \quad h = \pi/(n+1)$,

EV: $v^{(i)} = (v_1^{(i)}, v_2^{(i)}, ..., v_n^{(i)})^T, \quad v_j^{(i)} = \sin(ijh)$,

$$\lambda(A_3) \in \bigcup_{i=1}^{n} \mathcal{K}_i = \{z : |z-0| \le 1\} \cup \{z : |z-0| \le 2\} = [-2, 2].$$

7. Die EW von Diagonal- und Dreiecksmatrizen sind die Werte auf der Diagonalen der Matrix.
Für Blockdreiecksmatrizen

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} & \dots & A_{1m} \\ & A_{22} & A_{23} & \dots & A_{2m} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ 0 & & & & A_{mm} \end{pmatrix}$$

mit quadratischen Matrizen A_{ii} gilt

$$\det(A - \lambda I) = \det(A_{11} - \lambda I)\ \det(A_{22} - \lambda I) \cdot ... \cdot \det(A_{mm} - \lambda I).$$

8. Im Satz 2.15 haben wir die Ähnlichkeitstransformation einer beliebigen Matrix A auf die obere (komplexe) Dreiecksform $T = U^H AU = U^{-1}AU$ durchgeführt. Daraus ergeben sich Sonderfälle, die als eigenständige Sätze formuliert werden.

Satz 2.30 *Für eine hermitesche Matrix A existiert eine unitäre Matrix U ($U^H U = UU^H = I$), so dass die Matrix $T = U^H AU$ eine Diagonalmatrix ist mit den reellen Diagonalelementen t_{ii} als EW von A.*

Beweis. Bemerkung 2.3, 2. Teil.

Satz 2.31 Hauptachsentheorem, Spektralsatz

Ist A reell und symmetrisch, dann existiert eine Orthogonalmatrix Q ($Q^TQ = QQ^T = I$), so dass die Beziehung $Q^TAQ = Q^{-1}AQ = \Lambda = \mathrm{diag}(\lambda_i)$ erfüllt ist.
Die Spalten von Q sind wegen $AQ = Q\Lambda$ die orthonormalen EV von A.

Beweis. (Skizze)
Man führt den Nachweis mittels vollständiger Induktion bez. der Dimension n.
1. Sei $n = 1$.
Für $A = (a_{11})$ nimmt man $Q = I$ und somit $\Lambda = (\lambda_1) = (a_{11})$.

2. Die Gültigkeit für die Dimension $n-1$ bedeutet $P^TAP = \Lambda_{n-1},\ P^TP = I$.

3. Im Induktionsschritt für n betrachten wir die Matrix $A(n,n)$ mit dem EW λ_1 und zugehörigen EV $x_1,\ \|x_1\|_2 = 1$.
Man konstruiert nun $n-1$ Vektoren $y_1, y_2, ..., y_{n-1} \in \mathbb{R}^n$ unter Verwendung eines Orthogonalisierungsverfahrens mit folgender Eigenschaft:

$$\begin{aligned}(x_1, y_j) &= 0, \quad j = 1, 2, ..., n-1,\\ (y_i, y_j) &= \delta_{ij}, \quad i, j = 1, 2, ..., n-1.\end{aligned}$$

Damit gelten

$$\begin{aligned}Q_1^TQ_1 &= I, \quad Q_1 = (x_1, y_1, ..., y_{n-1}) = (x_1, Y),\\ Q_1^TAQ_1 &= \begin{pmatrix} \lambda_1 & 0^T \\ 0 & C \end{pmatrix}\\ &\quad \text{symm.}, \quad C = Y^TAY = C^T,\ 0 = \lambda_1 Y^Tx_1,\\ P^TCP &= \Lambda_{n-1}, \quad P^TP = I \quad \text{wegen Induktionsvoraussetzung},\\ Q_2 &= \begin{pmatrix} 1 & 0^T \\ 0 & P \end{pmatrix},\\ Q_2^TQ_1^TAQ_1Q_2 &= Q_2^T \begin{pmatrix} \lambda_1 & 0^T \\ 0 & C \end{pmatrix} Q_2,\\ &\quad \text{mit } Q = Q_1Q_2 \text{ als Orthogonalmatrix},\\ &= \begin{pmatrix} \lambda_1 & 0^T \\ 0 & P^TCP \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0^T \\ 0 & \Lambda_{n-1} \end{pmatrix} = \Lambda_n,\end{aligned}$$

und somit die Behauptung. □

9. Die Ähnlichkeitstransformation $T^{-1}AT$ einer reellen Matrix A mit ausschließlich reellen EW kann noch auf eine besondere Form gebracht werden. Dazu brauchen wir den Begriff der Jordanschen Normalform.

Definition 2.25 Jordansche Normalform

Eine Blockdiagonalmatrix $J = \mathrm{diag}(J_1, J_2, ..., J_p)$ *mit den quadratischen Jordan-Zellen (Jordan-Kästchen)*

$$J_j(t,t)(\lambda) = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda \end{pmatrix}, \quad j = 1, 2, ..., p. \tag{2.36}$$

der Dimension $t \times t$ *heißt Jordansche Normalform. Im Fall* $t = 1$ *vereinfacht sich die Jordan-Zelle zu* $J_j(1,1)(\lambda) = (\lambda)$.

Satz 2.32 *Jede reelle Matrix mit reellen EW ist ähnlich zu einer Matrix in Jordan-Normalform. Zu jedem EW* λ_i *mit seiner algebraischen und geometrischen Vielfachheit* n_i *bzw.* m_i *(*$1 \le m_i \le n_i \le n$*) gehören* m_i *Jordan-Zellen, wobei die Summe ihrer Dimensionen gleich* n_i *ist. Wenn wir* q *verschiedene EW* λ_i *voraussetzen, können wir die Jordan-Normalform als*

$$J = T^{-1}AT = \begin{pmatrix} J_1(k_1^{(1)}, k_1^{(1)})(\lambda_1') & & & & & \\ & \ddots & & & & \\ & & J_{m_1}(k_{m_1}^{(1)}, k_{m_1}^{(1)})(\lambda_1') & & & \\ & & & \ddots & & \\ & & & & J_1(k_1^{(q)}, k_1^{(q)})(\lambda_q') & \\ & & & & & \ddots \\ & & & & & J_{m_q}(k_{m_q}^{(q)}, k_{m_q}^{(q)})(\lambda_q') \end{pmatrix} \tag{2.37}$$

notieren, wobei

$$\begin{aligned} \lambda_1' &= \lambda_1 = \lambda_2 = ... = \lambda_{n_1}, \\ \lambda_2' &= \lambda_{n_1+1} = \lambda_{n_1+2} = ... = \lambda_{n_1+n_2}, \\ \lambda_3' &= \lambda_{n_1+n_2+1} = \lambda_{n_1+n_2+2} = ... = \lambda_{n_1+n_2+n_3}, \\ &\dots \\ \lambda_q' &= \lambda_{n_1+n_2+...+n_{q-1}+1} = \lambda_{n_1+n_2+...+n_{q-1}+2} = ... = \lambda_{n_1+n_2+...+n_q}, \quad n = \sum_{i=1}^{q} n_i, \end{aligned}$$

$$J_j(k_j^{(i)}, k_j^{(i)})(\lambda_i') = \begin{pmatrix} \lambda_i' & 1 & & & \\ & \lambda_i' & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i' \end{pmatrix}, \quad j = 1, 2, ..., m_i, \quad n_i = \sum_{j=1}^{m_i} k_j^{(i)}, \quad i = 1, 2, ..., q.$$

Beweis. [2], [11]

Der Satz gilt auch in Fall komplexer EW, wobei dann T komplexe Elemente enthält.

Folgerung 2.33 *Wenn in* $J = \mathrm{diag}(J_1, J_2, ..., J_p)$ *der obere Index* $p = n$ *ist, dann hat* J *die Form einer Diagonalmatrix, die Jordan-Zellen sind eindimensional, die Matrix* A *ist diagonalisierbar, ihre Elementarteiler* $\lambda_i - \lambda$ *sind ausschließlich linear und es gibt zu jedem EW* λ_i *ein System von* $m_i = n_i$ *linear unabhängigen EV. Damit ist das Gesamtsystem der* n *EV linear unabhängig.*

Beispiel 2.16

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}, \quad \text{EW}: \ \lambda_1 = 1, \ \lambda_2 = 2,$$

EV: $x_1 = (1, -1)^T, \ x_2 = (0, 1)^T$ linear unabhängig,

$$X = (x_1, x_2) = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \ X^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \ X^{-1}AX = \mathrm{diag}(\lambda_1, \lambda_2) = \mathrm{diag}(J_1, J_2).$$

Beispiel 2.17

Die Matrix $A(9, 9)$ möge folgende EW mit ihren Vielfachheiten haben:

$\lambda_1' = \ \ 1 = \lambda_1 = ... = \lambda_5, \ m_1 = 2, \ n_1 = 5,$

$\lambda_2' = -1 = \lambda_6 = ... = \lambda_9, \ m_1 = 3, \ n_2 = 4.$

Ihre Jordan-Normalform mit den $p = 5$ Jordan-Zellen ist

$$J = \begin{pmatrix} \boxed{\begin{matrix} 1 & 1 & \\ & 1 & 1 \\ & & 1 \end{matrix}} & & & & \\ & \boxed{\begin{matrix} 1 & 1 \\ & 1 \end{matrix}} & & & \\ & & \boxed{\begin{matrix} -1 & 1 \\ & -1 \end{matrix}} & & \\ & & & \boxed{-1} & \\ & & & & \boxed{-1} \end{pmatrix}.$$

Die Elementarteiler sind $(1-\lambda)^3, \ (1-\lambda)^2, \ (-1-\lambda)^2, \ (-1-\lambda), \ (-1-\lambda)$.
Das charakteristische Polynom hat die Gestalt $p_9(\lambda) = (-1)^9(\lambda - 1)^5(\lambda + 1)^4$.

Beispiel 2.18

$$A = A^T = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}, \quad \text{EW}: \ \lambda_1 = 2, \ \lambda_2 = 4,$$

EV: $x_1 = (1, 1)^T, \ x_2 = (-1, 1)^T$ orthogonal,

$$X = (x_1, x_2) = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \ \text{orthogonal}, \ X^{-1} = \tfrac{1}{2}X^T, \ \ A = X \ \mathrm{diag}(\lambda_1, \lambda_2) \ X^{-1}.$$

Man kann (x_1, x_2) aus einer Transformation des rechtwinkeligen Koordinatensystems

(e_1, e_2) ableiten und die Streckung der EV darstellen. Die EV der zu A ähnlichen Matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ sind die Einheitsvektoren.

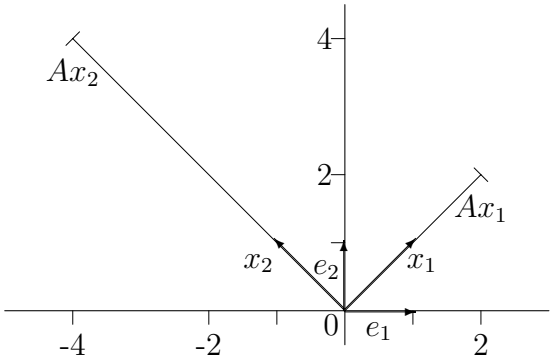


Abb. 2.1 Transformation $AX = X\Lambda$ im Koordinatensystem

Beispiel 2.19

$$A = A^T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \text{EW}: \ \lambda_1 = 3, \ \lambda_{2,3} = 0,$$

EV: $\tilde{x}_1 = (1,1,1)^T$, $\tilde{x}_2 = (-1,1,0)^T$, $\tilde{x}_3 = (-1,2,-1)^T$ lin. unabhängig, $\tilde{x}_1 \perp \tilde{x}_{2,3}$,
$x_1 = (1,1,1)^T$, $x_2 = (-1,1,0)^T$, $x_3 = (1,1,-2)^T$ orthogonal,

$$X = (x_1, x_2, x_3) = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & -2 \end{pmatrix} \quad \text{orthogonale Modalmatrix.}$$

Die algebraische Vielfachheit n_2 des EW λ_2 ist gleich dem Rangabfall m_2 der Matrix $A - \lambda_2 I$, also $m_2 = 3 - \mathrm{rang}(A - \lambda_2 I) = 3 - \mathrm{rang}(A) = 3 - 1 = 2$.
Bei symmetrischen Matrizen kann man also anstelle der n linear unabhängigen EV gleich mit orthogonalen EV arbeiten.

Beispiel 2.20

$$A = \begin{pmatrix} 5 & -4 \\ 1 & 1 \end{pmatrix}, \quad \text{EW}: \ \lambda_{1,2} = 3.$$

Es gibt nur einen EV $x_1 = (2,1)^T$. Der Rangabfall der Matrix $A - \lambda_1 I$ ist Eins.
Mit der Wahl eines zweiten Vektors y, der linear unabhängig von x_1 ist, erhält man

$$X = (x_1, y) = \begin{pmatrix} 2 & -1 \\ 1 & -1 \end{pmatrix}, \quad X^{-1} = \begin{pmatrix} 1 & -1 \\ 1 & -2 \end{pmatrix}, \quad X^{-1}AX = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix} = \mathrm{diag}(J_1),$$

und damit die gewünschte Jordan-Normalform.

10. Zwischen den EW, der Spur und der Frobenius-Norm einer Matrix erhält man den folgenden Zusammenhang.

Satz 2.34 *Für die EW der reellen Matrix $A = (a_{ij})$ gilt*

$$\sum_{i=1}^{n} |\lambda_i|^2 \;\leq\; \text{spur}(AA^T) \;=\; \text{spur}(A^TA) \;=\; \|A\|_F^2 \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2. \qquad (2.38)$$

Beweis. (Skizze)
Gemäß Satz 2.15 erhält man $T = U^HAU = U^{-1}AU = (t_{ij})$ obere Dreiecksmatrix mit den EW t_{ii} auf der Diagonalen, die wegen der Ähnlichkeit auch EW von A sind. Deshalb ist

$$\sum_{i=1}^{n} |\lambda_i(A)|^2 \;=\; \sum_{i=1}^{n} |\lambda_i(T)|^2 \;=\; \sum_{i=1}^{n} |t_{ii}|^2 \;\leq\; \sum_{i=1}^{n}\sum_{j=1}^{n} |t_{ij}|^2 \;=\; \text{spur}(TT^H).$$

Weiter sind wegen

$$TT^H \;=\; U^HAUU^HA^HU \;=\; U^H(AA^H)U$$

die Matrizen TT^H und AA^H ähnlich, haben damit die gleichen EW und wegen $\text{spur}(V) = \sum_{i=1}^{n} EW(V)$ die gleiche Spur. Somit ist

$$\sum_{i=1}^{n} |\lambda_i(A)|^2 \;\leq\; \text{spur}(TT^H) \;=\; \text{spur}(AA^T) \;=\; \text{spur}(A^TA) \;=\; \|A\|_F^2,$$

wobei die letzten beiden Gleichheiten einfach nachzurechnen sind. □

11. Die Matrix A hat das charakteristische Polynom

$$p_n(\lambda) \;=\; \det(A - \lambda I) \;=\; (-1)^n\lambda^n + c_1\lambda^{n-1} + ... + c_{n-1}\lambda + c_n.$$

Das gleiche charakteristische Polynom erhalten wir mit der einfachen Matrix

$$\hat{A}(n,n) = \begin{pmatrix} 0 & 1 & . & \cdots & . & 0 \\ 0 & 0 & 1 & \cdots & . & . \\ 0 & 0 & 0 & \cdots & . & . \\ . & . & . & \cdots & . & . \\ . & . & . & \cdots & 1 & . \\ . & . & . & \cdots & 0 & 1 \\ (-1)^{n+1}c_n & (-1)^{n+1}c_{n-1} & (-1)^{n+1}c_{n-2} & \cdots & (-1)^{n+1}c_2 & (-1)^{n+1}c_1 \end{pmatrix}.$$

Die Matrix $\hat{A}$ wird auch als **Begleitmatrix** oder erzeugende Matrix bezeichnet. Den Nachweis kann man führen, indem man die Determinante $\det(\hat{A} - \lambda I)$ nach der ersten Spalte oder auch letzten Zeile entwickelt.

2.2.2 Eigenschaften von Matrizen in Bezug auf Eigenwerte

1. Hat die Matrix A den EW λ, so sind λ^m EW von A^m und $1 - c\lambda^m,\ c \in \mathbb{C}$, EW von $I - cA$.

2. Wenn die Matrix A hermitesch ist, dann sind alle ihre EW $\lambda(A)$ reell. Insbesondere gilt das für symmetrische Matrizen.
Diese Aussage ist auch das Ergebnis von Satz 2.3 oder Satz 2.15 und Bemerkung 2.3.

3. Sei $A = A^H$. Dann gilt: A ist positiv definit gdw. $\lambda(A) > 0$ ist (Satz 2.16).

4. Die Matrizen $A^T A$ und AA^T sind symmetrisch positiv semidefinit und haben nur reelle nicht negative EW. Beide haben das gleiche Spektrum.

Beweis.
Die Symmetrie gilt wegen $(A^T A)^T = A^T (A^T)^T = A^T A$, analog für AA^T. Im Ergebnis von Satz 2.3 sind die EW reell.
Die positive Semidefinitheit folgt aus $x^T A^T A x = (y^T y) = \sum_{i=1}^{n} y_i^2 \geq 0$, analog ist $AA^T \geq 0$. Nach Satz 2.16 sind ihre EW nicht negativ.
Zu zeigen bleibt, dass $A^T A$ und AA^T die gleichen EW haben.
(1) Wenn A regulär ist, dann gilt die Behauptung wegen der Ähnlichkeit beider Matrizen gemäß

$$A^T A = A^{-1} A\, A^T A = A^{-1} (AA^T) A.$$

(2) Im allgemeinen Fall zeigen wir, dass die positiven EW von $A^T A$ genau auch die von AA^T sind, was dann zugleich bedeutet, dass sie die restlichen Null-EW ebenfalls gemeinsam haben.
Sei $A^T A x = \lambda x,\ x \neq 0$, und $\lambda > 0$ einer der Nichtnull-EW von $A^T A$.
Die Multiplikation der Gleichung mit A von links ergibt

$$AA^T Ax = \lambda Ax,\ y = Ax, \quad \rightarrow \quad AA^T y = \lambda y.$$

Es ist also $y = Ax$ ein EV von AA^T mit demselben EW λ, denn $y = Ax \neq 0$ wegen $A^T Ax \neq 0$. Genauso kann man für $AA^T y = \lambda y,\ y \neq 0$ argumentieren. □

In [30] findet man einen allgemeineren Satz für rechteckige komplexe Matrizen.
Für eine symmetrische Matrix ist die Aussage trivial.

Beispiel 2.21 Gleichheit von $\sigma(A^T A) = \sigma(AA^T)$
Sei

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

Damit sind

$$A^T = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad A^T A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad AA^T = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix},$$

die charakteristischen Polynome $\det(A^TA-\lambda I)=(1-\lambda)^2-1$ und $\det(AA^T-\lambda I)=(2-\lambda)(-\lambda)$ identisch und das gemeinsame Spektrum $\sigma=\{2,0\}$.

5. Seien $\lambda_1,\lambda_2,...,\lambda_s$ paarweise verschiedene EW von A. Dann sind die zugehörigen EV linear unabhängig.

Beweis. Der Nachweis erfolgt mittels vollständiger Induktion und Anwendung der Definition der linearen Unabhängigkeit von Vektoren.

(1) Seien $\lambda_1\neq\lambda_2$ und $x_{1,2}\neq 0$ die zugehörigen EV. Es folgt

$$\begin{aligned}
0 &= c_1x_1+c_2x_2,\\
0 &= A(c_1x_1+c_2x_2)\\
&= c_1\lambda_1x_1+c_2\lambda_2x_2,\\
0 &= c_1\lambda_1x_1+c_2\lambda_1x_2,\ \lambda_1\neq 0 \text{ bzw. } \lambda_2\neq 0,\\
0 &= c_2(\lambda_2-\lambda_1)x_2 \text{ durch Subtraktion der letzten beiden Zeilen,}\\
&\Rightarrow\ c_2=0\ \Rightarrow\ c_1=0\\
&\Rightarrow\ x_{1,2} \text{ linear unabhängig.}
\end{aligned}$$

(2) Seien $\lambda_i,\ i=1,2,...,k$, paarweise verschieden und die zugehörigen EV $x_i\neq 0$ linear unabhängig. Nehmen wir den nächsten EV x_{k+1} hinzu.

$$\begin{aligned}
0 &= c_1x_1+c_2x_2+...+c_kx_k+c_{k+1}x_{k+1},\\
0 &= A(c_1x_1+c_2x_2+...+c_kx_k+c_{k+1}x_{k+1})\\
&= c_1\lambda_1x_1+c_2\lambda_2x_2+...+c_k\lambda_kx_k+c_{k+1}\lambda_{k+1}x_{k+1}.
\end{aligned}$$

Wir führen eine Fallunterscheidung durch.

1. $\lambda_{k+1}=0\ \Rightarrow\ 0=c_1\lambda_1x_1+c_2\lambda_2x_2+...+c_k\lambda_kx_k$

$$\begin{aligned}
&\Rightarrow\ c_1=...=c_k=0\\
&\Rightarrow\ c_{k+1}=0.
\end{aligned}$$

2. $\lambda_{k+1}\neq 0\ \Rightarrow\ 0=c_1\lambda_{k+1}x_1+c_2\lambda_{k+1}x_2+...+c_k\lambda_{k+1}x_k+c_{k+1}\lambda_{k+1}x_{k+1},$

$$\begin{aligned}
0=c_1(\lambda_1-\lambda_{k+1})x_1+c_2(\lambda_2-\lambda_{k+1})x_2+...+\\
c_k(\lambda_k-\lambda_{k+1})x_k+c_{k+1}(\lambda_{k+1}-\lambda_{k+1})x_{k+1}
\end{aligned}$$

d. h.

$$\begin{aligned}
0 &= c_1(\lambda_1-\lambda_{k+1})x_1+c_2(\lambda_2-\lambda_{k+1})x_2+...+c_k(\lambda_k-\lambda_{k+1})x_k\\
&\Rightarrow\ c_1=...=c_k=0\\
&\Rightarrow\ c_{k+1}=0.
\end{aligned}$$

Damit sind die Vektoren $x_1,...,x_k,x_{k+1}$ linear unabhängig. □

6. Sei $A = A^T$ reell. Die EV zu paarweise verschiedenen EW sind orthogonal.

Beweis. (Skizze) Es sei $\lambda_1 \neq \lambda_2$ mit den EV x_1, x_2.

$$\begin{aligned} x_2^T A x_1 &= x_2^T \lambda_1 x_1 = \lambda_1 x_2^T x_1, \\ x_2^T A^T x_1 &= (Ax_2)^T x_1 = \lambda_2 x_2^T x_1, \quad A = A^T, \\ 0 &= (\lambda_1 - \lambda_2) x_2^T x_1, \\ 0 &= x_2^T x_1. \end{aligned}$$

□

7. Seien alle EW λ_i von A reell und paarweise verschieden.
Dann bilden die EV $x_1, x_2, ..., x_n$ von A sowie die EV $y_1, y_2, ..., y_n$ von A^T zwei Basen des Raums $\mathbb{R}^n$, wobei sie biorthonormal sind, d. h.

$$(x_j, y_k) = x_j^T y_k = \delta_{jk}.$$

Beweis. (Skizze)
Es wurden schon im Punkt 5 sowie in der Folgerung 2.29 (1) gezeigt:
(a) Die EV x_i sind linear unabhängig und bilden somit eine Basis.
(b) Die EW von A^T sind dieselben wie von A, somit auch paarweise verschieden.
Die EV y_k von A^T genügen der Gleichung $A^T y_k = \lambda_k y_k$ und sind linear unabhängig.
Nun zeigt man die Biorthogonalität.

$$\begin{aligned} (Ax_j, y_k) &= (\lambda_j x_j, y_k) = \bar{\lambda}_j (x_j, y_k) = \lambda_j (x_j, y_k), \\ (Ax_j, y_k) &= (x_j, A^T y_k) = (x_j, \lambda_k y_k) = \lambda_k (x_j, y_k), \\ 0 &= (\lambda_j - \lambda_k)(x_j, y_k) \quad \Rightarrow \quad (x_j, y_k) = 0 \text{ für } j \neq k. \end{aligned}$$

Die Vektoren x_j, y_k können so gewählt werden, dass $(x_j, y_j) = 1$ ist.
Da beides Basen sind, gilt das Folgende.

$$\begin{aligned} x_j &= c_1 y_1 + c_2 y_2 + ... + c_n y_n, \\ (x_j, x_j) &= c_1 (x_j, y_1) + c_2 (x_j, y_2) + ... + c_n (x_j, y_n), \\ 0 < (x_j, x_j) &= c_j (x_j, y_j), \\ \alpha_j &= (x_j, y_j) \neq 0, \quad j = 1, 2, ..., n. \end{aligned}$$

Setzt man anstelle der Basis $\{y_k\}$ die Basis (und EV) $\{\frac{1}{\alpha_k} y_k\}$, so erhält man unmittelbar die Biorthonormalität. □

Folgerung 2.35 *Ist A symmetrisch mit paarweise verschiedenen EW, so kann man bez. der EV $y_j = x_j$, $j = 1, 2, ..., n$, setzen, wobei die x_j auf Eins normiert sind. Die EV x_j bilden ein ONS.*

8. Ist der EW λ einer reellen Matrix A komplex, so sind die Komponenten des EV x ebenfalls komplex. Wegen $A\overline{x} = \overline{Ax} = \overline{\lambda x} = \overline{\lambda}\overline{x}$ hat die Matrix A auch den konjugiert komplexen EW $\overline{\lambda}$ mit dem zugehörigen EV $\overline{x}$.

9. Diagonalähnliche Matrizen.
Falls bezüglich der Vielfachheiten $m_i = n_i$ für alle EW λ_i gilt, so heißt A diagonalähnlich. Die Gesamtheit aller n EV spannt dann den Raum $\mathbb{R}^n$ auf.

10. Die EW von Diagonal-/Dreiecksmatrizen sind ihre Elemente auf der Diagonalen.

11. Das charakterische Polynom einer Tridiagonalmatrix ist

$$p_n(\lambda) \;=\; \det \begin{pmatrix} a_1-\lambda & b_1 & 0 & \dots & & & 0 \\ c_1 & a_2-\lambda & b_2 & 0 & \dots & & \\ 0 & c_2 & a_3-\lambda & b_3 & 0 & \dots & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & & \dots & 0 & c_{n-2} & a_{n-1}-\lambda \quad b_{n-1} \\ 0 & & & & \dots & 0 & c_{n-1} \quad a_n-\lambda \end{pmatrix}.$$

Seine Notation in der Normalform sei

$$F_n \;=\; p_n(\lambda) \;=\; (-1)^n(d_n\lambda^n + d_{n-1}\lambda^{n-1} + ... + d_1\lambda + d_0), \quad d_n = 1.$$

Die Berechnung des Funktionswerts im Zusammenhang mit der Lösung der Nullstellenaufgabe $p_n(\lambda) = 0$ erfolgt über die Entwicklung der Determinante nach der letzten Spalte und dann nach der letzten Zeile. Das führt auf die rekursive Formel (Drei-Term-Rekursion)

$$F_k \;=\; (a_k - \lambda)F_{k-1} - b_{k-1}c_{k-1}F_{k-2}, \; k = 2, 3, ..., n, \quad F_0 \;=\; 1, \quad F_1 \;=\; a_1 - \lambda.$$

2.2.3 Eigenschaften von Eigenvektoren

Fassen wir einige wichtige Eigenschaften noch einmal zusammen.

1. Das EWP $Ax = \lambda x$ notiert man in Matrixform als $AX = X\Lambda$, wobei
$\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_r)$ und die zugehörigen linear unabhängigen EV $x_i, \; i = 1, 2, ..., r$, $r \leq n$, die Spalten der Modalmatrix X bilden.

2. Die EV sind bis auf einen Proportionalitätsfaktor festgelegt.
Ein EV $x \in \mathbb{R}^n$ mit $\|x\|_2 = 1$ heißt normierter EV.

3. Die EV x_i, x_j zu verschiedenen EW λ_i, λ_j sind linear unabhängig.

4. Ist $A = A^T$, so sind die EV x_i, x_j zu verschiedenen EW λ_i, λ_j orthogonal.

5. Zum EW $\lambda_i, \; i = 1, 2, ..., s$, mit der algebraischen Vielfachheit n_i existieren genau m_i linear unabhängige EV, die damit einen m_i-dimensionalen Unterraum des $\mathbb{R}^n$ aufspannen (geometrische Vielfachheit). Dabei ist stets $1 \leq m_i \leq n_i, \; i = 1, 2, ..., s$.

Zum EW λ_i mit der geometrischen Vielfachheit m_i kann man die Matrixdarstellung $AX = X\Lambda = X\lambda_i$ notieren, wobei $\Lambda = \Lambda(m_i, m_i) = \text{diag}(\lambda_i, \lambda_i, ..., \lambda_i)$ und die Modalmatrix $X = X(n, m_i)$ die m_i zugehörigen linear unabhängigen EV $x_i^{(j)},\ j = 1, 2, ..., m_i$, als Spalten enthält. Diese Vektoren sind eine Basis X_i des linearen Unterraums $[\mathrm{X}_i] = \text{span}\,\mathrm{X}_i \subset \mathbb{R}^n$, auch Eigenraum genannt.
Damit ist A nicht nur allgemein eine Abbildung $\mathbb{R}^n \to \mathbb{R}^n$, sondern es gilt insbesondere $A:\ [\mathrm{X}_i] \to [\mathrm{X}_i]$, was die Invarianz der Eigenräume bedeutet.
Zum EW λ_i mit seinen EV $x_i^{(j)},\ j = 1, 2, ..., m_i$, ist auch jede nicht verschwindende Linearkombination dieser ein EV. Somit kann der Eigenraum zu λ_i aus orthogonalen oder orthonormalen EV gebildet werden.
Ist $A = A^T$, so gilt $m_i = n_i,\ i = 1, 2, ..., s$, und man wählt alle EV orthogonal. Alle EV zusammen genommen, d. h. alle Basen $\mathrm{X}_i = \{x_i^{(j)}\}_{j=1}^{m_i},\ i = 1, 2, ..., s$, erhält man die direkte Summe der Unterräume $[\mathrm{X}_1] \oplus [\mathrm{X}_2] \oplus ... \oplus [\mathrm{X}_s] = \mathbb{R}^n$.

6. Die Situation vereinfacht sich, wenn wie in der Folgerung 2.33 die Matrix A diagonalisierbar ist, d. h. $J = T^{-1}AT$ ist eine Diagonalmatrix der EW und die Spalten t_i von T sind die n zugehörigen linear unabhängigen EV. Für eine symmetrische Matrix ist das auch der Fall.
Gleichzeitig haben dann wir nach Satz 2.31 (Hauptachsentheorem) die Beziehung $Q^TAQ = Q^{-1}AQ = \Lambda = \text{diag}(\lambda_i)$ mit einer Orthogonalmatrix Q, so dass wegen $AQ = Q\Lambda$ die Spalten q_i von Q die orthonormalen EV von A sind. $\{q_i\}_{i=1}^n$ ist ein ONS.
Zwischen den beiden Basen $\mathrm{T} = \{t_i\}_{i=1}^n$ und $\mathrm{Q} = \{q_i\}_{i=1}^n$ des $\mathbb{R}^n$ existiert eine Basistransformation mit einer eindeutigen regulären Abbildungsmatrix $S \in \mathbb{R}^{n,n}$, so dass $Q = TS$ ist.
Wir haben also die Beziehungen

$$\begin{aligned} AT &= TJ = T\Lambda, \quad A = A^T \text{ oder } A \text{ diagonalisierbar}, \\ AQ &= Q\Lambda, \quad Q = TS, \\ ATS &= TS\Lambda, \\ AT &= TS\Lambda S^{-1}, \end{aligned}$$

woraus $\Lambda = S\Lambda S^{-1}$ bzw. $\Lambda S = S\Lambda$ folgt.

7. Wir wissen, dass bei Matrizen die Kommutativität i. Allg. nicht erfüllt ist.
Einige Aussagen zu ihrer Vertauschbarkeit macht der folgende Satz.

Satz 2.36 Vertauschbarkeit von Matrizen
Seien $A, B \in \mathbb{R}^{n,n}$.
(1) Falls A und B die gleiche Basis aus n orthonormalen EV besitzen, dann gilt $AB = BA$.
(2) Falls $AB = BA$ ist und von A die EV zu paarweise verschiedenen EW orthonormal sind, dann besitzen A und B die gleiche Basis aus orthonormalen EV.

Beweis.
Zu (1): Sei $Av_i = \lambda_i v_i$, $AV = V\Lambda_A$, analog $Bw_i = \mu_i w_i$, $BW = W\Lambda_B$, und

$$\mathrm{V} = \mathrm{W} = \mathrm{U} = \{u_1, u_2, ..., u_n\}, \quad U = \begin{pmatrix} | & | & & | \\ u_1 & u_2 & ... & u_n \\ | & | & & | \end{pmatrix}.$$

Dann folgt wegen der Vertauschbarkeit der Diagonalmatrizen Λ_A und Λ_B

$$\begin{aligned} ABU &= ABW = AW\Lambda_B = AV\Lambda_B = V\Lambda_A\Lambda_B \\ &= W\Lambda_B\Lambda_A = BW\Lambda_A = BV\Lambda_A = BAV = BAU, \quad \exists U^{-1}, \\ AB &= BA. \end{aligned}$$

Zu (2): Seien nun $AB = BA$ und $AV = V\Lambda_A$, sowie $\lambda_1, ..., \lambda_r$ die paarweise verschiedenen EW von A mit ihren algebraischen Vielfachheiten $n_1, ..., n_r$ und $\sum_{i=1}^{r} n_i = n$. Die geometrischen Vielfachheiten sind $m_i \leq n_i$.
Für jeden EW λ_i haben wir die m_i Beziehungen $Ax_i^{(j)} = \lambda_i x_i^{(j)}$, $j = 1, 2, ..., m_i$, mit den linear unabhängigen EV $x_i^{(j)}$. Den Eigenraum $[\mathrm{X}_i] = [\{x_i^{(1)}, ..., x_i^{(m_i)}\}]$ kann man auch durch ein ONS von m_i EV $u_i^{(j)}$ erzeugen, so dass

$$A \begin{pmatrix} | & & | \\ u_i^{(1)} & ... & u_i^{(m_i)} \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ u_i^{(1)} & ... & u_i^{(m_i)} \\ | & & | \end{pmatrix} \mathrm{diag}(\lambda_i, ..., \lambda_i).$$

Sei $\mathrm{U} = (\mathrm{U}_1, \mathrm{U}_2, ..., \mathrm{U}_r)$ eine Zerlegung des ONS von EV bez. der r Eigenräume.
Dann folgt $AU_i = U_i\Lambda_i = U_i\lambda_i$, $i = 1, 2, ..., r$. Hieraus ergibt sich

$$A\,BU_i = BAU_i = B(U_i\lambda_i) = (BU_i)\,\lambda_i.$$

Damit ist BU_i EV von A zum EW λ_i, d. h. $B : [\mathrm{U}_i] \to [\mathrm{U}_i]$ für jedes $i = 1, 2, ..., r$. B lässt also die Eigenräume von A invariant.
Seien nun $\mu_i^{(j)}$, $j = 1, 2, ..., m_i$, die EW von B bez. des Unterraums $[\mathrm{U}_i]$ und V_i ein ONS aus EV von B. Dann folgt

$$\begin{aligned} BV_i &= V_i\,\mathrm{diag}(\mu_i^{(1)}, \mu_i^{(2)}, ..., \mu_i^{(m_i)}) \;=\; V_i\Lambda_i, \\ AV_i &= V_i\lambda_i, \end{aligned}$$

da V_i nur eine andere Basis von $[\mathrm{U}_i]$ ist. Mit $\mathrm{V} = (\mathrm{V}_1, \mathrm{V}_2, ..., \mathrm{V}_r)$ als eine neue Basis folgt nun

$$\begin{aligned} AV &= V\Lambda_A, \\ BV &= V\Lambda_B, \end{aligned}$$

wobei $\Lambda_B = \mathrm{diag}(\Lambda_1, \Lambda_2, ..., \Lambda_r)$ ist.
D. h. beide Matrizen besitzen eine gemeinsame Basis V aus orthonormalen EV. □

Folgerung 2.37

(1) Für symmetrische Matrizen A und B vereinfacht sich der Satz zu:
Es gilt $AB = BA$ gdw. A und B die gleiche Basis aus orthonormalen EV besitzen.

(2) Die symmetrischen Matrizen A_k, $k = 1, 2, ..., m$, sind genau dann im Matrixprodukt paarweise vertauschbar, wenn es ein gemeinsames ONS aus EV gibt.

Beispiel 2.22

(1)
$$A = \begin{pmatrix} 5 & -4 \\ 1 & 1 \end{pmatrix}, \quad \text{EW}: \ \lambda_{1,2} = 3.$$

Es gibt nur einen (orthonormalen) EV $x_1 = (2/\sqrt{5}, 1/\sqrt{5})^T$. Der Rangabfall zur Matrix $A - \lambda_1 I$ ist 1.
Die Matrix
$$B = \begin{pmatrix} 6 & -4 \\ 1 & 2 \end{pmatrix}, \quad \text{EW}: \ \lambda_{1,2} = 4,$$
hat dasselbe ONS von EV, so dass
$$AB = BA = \begin{pmatrix} 26 & -28 \\ 7 & -2 \end{pmatrix}.$$

(2)
$$A(n,n) = \begin{pmatrix} 1 & 1 & & & \\ & 1 & 1 & & \\ & & \ddots & \ddots & \\ & & & 1 & 1 \\ & & & & 1 \end{pmatrix}, \quad \text{EW}: \ \lambda_{1,...,n} = 1.$$

Der Rang der Matrix $A - \lambda_1 I$ beträgt $n - 1$, so dass der Rangabfall 1 ist und das ONS von EV aus einem einzigen EV $x_1 = (1, 0, ..., 0)^T$ besteht. Den gleichen EV hat die Matrix
$$B(n,n) = \begin{pmatrix} 2 & 1 & 1 & \dots & 1 \\ & 2 & 1 & \dots & 1 \\ & & \ddots & \ddots & \vdots \\ & & & 2 & 1 \\ & & & & 2 \end{pmatrix}, \quad \text{EW}: \ \lambda_{1,...,n} = 2.$$

Es gilt
$$AB = BA = \begin{pmatrix} 2 & 3 & 2 & 2 & \dots & 2 \\ & 2 & 3 & 2 & \dots & 2 \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & 2 & 3 & 2 \\ & & & & 2 & 3 \\ & & & & & 2 \end{pmatrix}.$$

2.2.4 EWP und Matrixzerlegung

In der Defintion 2.23 haben wir die spezielle Zerlegung $A = D - E - F = D - C = D(I - L - U)$ der reellen Matrix A in ihre Diagonal- und Dreieckskomponenten erläutert. Diese soll die Grundlage für einige weitere Betrachtungen sein.

Definition 2.26 Konsistent geordnete Matrix
Eine Matrix mit ihrer Zerlegung $A = D(I - L - U)$ gemäß (2.16) heißt konsistent geordnet, falls die EW der Matrix

$$J(\alpha) \;=\; \alpha L + \frac{1}{\alpha} U, \tag{2.39}$$

wobei $\alpha \neq 0$ ist, unabhängig von α sind.

Satz 2.38 *Jede Blockmatrix der Form*

$$\left(\begin{array}{c|c} I_1 & A_{12} \\ \hline A_{21} & I_2 \end{array}\right), \quad I_k \text{ Einheitsmatrizen}, \tag{2.40}$$

jede Tridiagonalmatrix mit nicht verschwindenden Diagonalelementen sowie jede tridiagonale Blockmatrix der Gestalt

$$\begin{pmatrix} D_1 & A_{12} & & & 0 \\ A_{21} & D_2 & A_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & A_{m-1,m-2} & D_{m-1} & A_{m-1,m} \\ 0 & & & A_{m,m-1} & D_m \end{pmatrix} \tag{2.41}$$

mit quadratischen regulären Diagonalmatrizen D_i ist konsistent geordnet.

Beweis.
Wir zeigen für jede der genannten Matrizen die Ähnlichkeit mir einer Matrix, deren EW nicht vom Parameter α abhängen.
(1) (2×2)-Blockmatrix

$$A = \begin{pmatrix} I_1 & A_{12} \\ A_{21} & I_2 \end{pmatrix} = I - \begin{pmatrix} 0 & 0 \\ -A_{21} & 0 \end{pmatrix} - \begin{pmatrix} 0 & -A_{12} \\ 0 & 0 \end{pmatrix} = I(I - L - U).$$

$$\begin{aligned} J(\alpha) &= \alpha L + \alpha^{-1} U \\ &= \begin{pmatrix} 0 & -\alpha^{-1} A_{12} \\ -\alpha A_{21} & 0 \end{pmatrix} \\ &= \begin{pmatrix} I_1 & 0 \\ 0 & \alpha I_2 \end{pmatrix} \begin{pmatrix} 0 & -A_{12} \\ -A_{21} & 0 \end{pmatrix} \begin{pmatrix} I_1 & 0 \\ 0 & \alpha^{-1} I_2 \end{pmatrix} \\ &= S(\alpha) J(1) S(\alpha)^{-1}, \quad S(\alpha) = \operatorname{diag}(I_1, \alpha I_2). \end{aligned}$$

(2) Tridiagonalmatrix

$$A = \begin{pmatrix} a_{11} & a_{12} & & & 0 \\ a_{21} & a_{22} & a_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & & & a_{n,n-1} & a_{n,n} \end{pmatrix} = D(I - L - U),$$
$$D = \mathrm{diag}(A), \quad a_{ii} \neq 0,$$

$$J = L + U = -\begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & & & 0 \\ \frac{a_{21}}{a_{22}} & 0 & \frac{a_{23}}{a_{22}} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{a_{n-1,n-2}}{a_{n-1,n-1}} & 0 & \frac{a_{n-1,n}}{a_{n-1,n-1}} \\ 0 & & & \frac{a_{n,n-1}}{a_{n,n}} & 0 \end{pmatrix} = 1 \cdot L + 1^{-1} \cdot U = J(1),$$

$$J(\alpha) = \alpha L + \alpha^{-1} U = S(\alpha) J(1) S(\alpha)^{-1}, \quad S(\alpha) = \mathrm{diag}(1, \alpha, \alpha^2, ..., \alpha^{n-1}).$$

(3) Blocktridiagonalmatrix

$$A = \begin{pmatrix} D_1 & A_{12} & & & 0 \\ A_{21} & D_2 & A_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & A_{m-1,m-2} & D_{m-1} & A_{m-1,m} \\ 0 & & & A_{m,m-1} & D_m \end{pmatrix} = D(I - L - U),$$
$$D = \mathrm{diag}(D_1, ..., D_m), \ \exists\, D_k^{-1},$$

$$J = L+U = -\begin{pmatrix} 0 & D_1^{-1}A_{11} & & & 0 \\ D_2^{-1}A_{21} & 0 & D_2^{-1}A_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & D_{m-1}^{-1}A_{n-1,n-2} & 0 & D_{m-1}^{-1}A_{n-1,n} \\ 0 & & & D_m^{-1}A_{n,n-1} & 0 \end{pmatrix} = J(1),$$

$$J(\alpha) = \alpha L + \alpha^{-1} U = S(\alpha) J(1) S(\alpha)^{-1}, \quad S(\alpha) = \mathrm{diag}(I_1, \alpha I_2, \alpha^2 I_3, ..., \alpha^{m-1} I_m).$$

□

Bemerkung 2.4 Die Tridiagonalität der Matrix A ist in diesem Satz eine wichtige Voraussetzung.

Natürlich setzt die Notation der Zerlegung $A = D(I - L - U)$ stillschweigend die Regularität der Matrix D voraus. Was macht man aber z. B. im Fall

$$\left(\begin{array}{c|c} O_1 & A_{12} \\ \hline A_{21} & O_2 \end{array}\right), \quad O_k \text{ quadratische Nullmatrizen?}$$

Dann ist

$$A = -L - U, \quad L = \begin{pmatrix} 0 & 0 \\ -A_{21} & 0 \end{pmatrix}, \; U = \begin{pmatrix} 0 & -A_{12} \\ 0 & 0 \end{pmatrix}, \quad J = L + U.$$

Wir zeigen nun auf einem anderen Weg, dass die Matrizen $J(\alpha) = \alpha L + \alpha^{-1} U$ und $J = J(1)$ die gleichen EW haben, d. h. es gilt $Jx = \lambda x$ gdw. $J(\alpha)y = \lambda y$ mit $x, y \neq 0$. Entsprechend der Blockstruktur von J setzt sich der Vektor x aus den beiden Teilvektoren u, v zusammen, so dass gilt

$$Jx = \begin{pmatrix} 0 & -A_{12} \\ -A_{21} & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -A_{12}v \\ -A_{21}u \end{pmatrix} = \lambda \begin{pmatrix} u \\ v \end{pmatrix} = \lambda x.$$

Wir brauchen nur zu überprüfen, ob der Vektor y mit den Teilvektoren $u, \alpha v$ dem EWP $J(\alpha)y = \lambda y$ genügt.

$$\begin{aligned} J(\alpha)y &= \begin{pmatrix} 0 & -\alpha^{-1}A_{12} \\ -\alpha A_{21} & 0 \end{pmatrix} \begin{pmatrix} u \\ \alpha v \end{pmatrix} \\ &= \begin{pmatrix} -A_{12}v \\ -\alpha A_{21}u \end{pmatrix} = \begin{pmatrix} \lambda u \\ \alpha \lambda v \end{pmatrix} = \lambda \begin{pmatrix} u \\ \alpha v \end{pmatrix} = \lambda y. \end{aligned}$$

Ergänzend betrachten wir noch Beispielmatrizen, die durchaus gemäß Satz 2.38 konsistent geordnet sind, aber bei denen wir die entsprechende Kontrolle einfach mit der Berechnung der EW machen wollen.

Beispiel 2.23 Konsistent geordnete tridiagonale Matrix

$$A = \begin{pmatrix} 1 & b & 0 \\ a & 1 & d \\ 0 & c & 1 \end{pmatrix}, \quad J = L + U = \begin{pmatrix} 0 & 0 & 0 \\ -a & 0 & 0 \\ 0 & -c & 0 \end{pmatrix} + \begin{pmatrix} 0 & -b & 0 \\ 0 & 0 & -d \\ 0 & 0 & 0 \end{pmatrix},$$

$$J(\alpha) = \alpha L + \alpha^{-1} U.$$

Die EW sind unabhängig von α wegen

$$0 = \det(J(\alpha) - \lambda I) = \det \begin{pmatrix} -\lambda & -\alpha^{-1}b & 0 \\ -\alpha a & -\lambda & -\alpha^{-1}d \\ 0 & -\alpha c & -\lambda \end{pmatrix} = \lambda(ab + dc - \lambda^2).$$

Beispiel 2.24 Konsistent geordnete untere Dreiecksmatrix

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad J = L + U = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ -1 & -1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$J(\alpha) = \alpha L + \alpha^{-1} U.$$

Die EW sind unabhängig von α wegen

$$0 = \det(J(\alpha) - \lambda I) = \det \begin{pmatrix} -\lambda & 0 & 0 \\ -\alpha & -\lambda & 0 \\ -\alpha & -\alpha & -\lambda \end{pmatrix} = -\lambda^3.$$

Beispiel 2.25 Nicht konsistent geordnet ist die voll besetzte, symmetrische und positiv semidefinite Matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = I - L - U = I - J, \quad \sigma(A) = \{0, 0, 3\}.$$

$$J(\alpha) = \alpha L + \alpha^{-1} U = \begin{pmatrix} 0 & -\alpha^{-1} & -\alpha^{-1} \\ -\alpha & 0 & -\alpha^{-1} \\ -\alpha & -\alpha & 0 \end{pmatrix}$$

und ihre EW sind Nullstellen der charakteristischen Gleichung

$$0 = \det(J(\alpha) - \lambda I) = -\lambda^3 + 3\lambda - (\alpha + \alpha^{-1}),$$

die abhängig von α sind. Zu $\alpha = 1$ ergeben sich für $J = J(1) = I - A$ die EW $1, 1, -2$.

Analog ist die Situation für die spd Matrix

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix} = I - L - U = I - J, \; 0 < a < 1, \quad \sigma(A) = \{1 - a, 1 - a, 1 + 2a\}.$$

$$J(\alpha) = \alpha L + \alpha^{-1} U = \begin{pmatrix} 0 & -\alpha^{-1} a & -\alpha^{-1} a \\ -\alpha a & 0 & -\alpha^{-1} a \\ -\alpha a & -\alpha a & 0 \end{pmatrix}$$

und ihre EW sind Nullstellen der charakteristischen Gleichung

$$0 = \det(J(\alpha) - \lambda I) = -\lambda^3 + 3\lambda - (\alpha + \alpha^{-1}),$$

die abhängig von α sind.

Beispiel 2.26 Nicht konsistent geordnet ist die Tridiagonalmatrix

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Für A ist es wegen gleichzeitig verschwindender Elemente und NNE auf der Diagonalen einfach nicht möglich, eine der Darstellungen als Zerlegung $A = D(I - L - U), \; D = \mathrm{diag}(A)$ regulär, bzw. $A = -L - U, \; D = \mathrm{diag}(A) = 0$, anzugeben.

Die konsistente Ordnung einer Matrix A hat Auswirkung auf Eigenschaften der abgeleiteten Matrix J und damit im Weiteren auf die Konvergenz von IV zur Lösung des LGS $Ax = b$.

Satz 2.39 *Wenn die Matrix A positiv definit und mit der regulären Zerlegung $A = D - C = D(I - L - U),\ D = \mathrm{diag}(A)$, konsistent geordnet ist, dann ist mit μ als EW von $J = D^{-1}C = L + U$ auch die Größe $-\mu$ EW von J.*

Beweis. Aus $A = (a_{ij}) > 0$ folgen $a_{ii} > 0$ für alle i und $D > 0$.
Da A konsistent geordnet ist, sind die EW von $J(\alpha) = \alpha L + \alpha^{-1}U,\ \alpha \neq 0$, unabhängig von α.
Sei μ EW von $J = J(1) = L + U$. Dann ist μ auch EW von $J(-1) = -L - U$, d. h. aber

$$0 = \det(J(-1) - \mu I) = \det(-J - \mu I) \quad \Rightarrow \quad 0 = \det(J + \mu I) = \det(J - (-\mu)I)$$

und damit ist $-\mu$ EW von J. □

Den Teil $A > 0$ der Voraussetzung dieses Satzes kann man abschwächen zu
$D = \mathrm{diag}(A)$ regulär bzw. $a_{ii} \neq 0\ \forall\ i$.

Nicht immer kann man so einfach wie im Satz 2.38 feststellen, ob eine Matrix konsistent geordnet ist. Manchmal macht man sich deshalb andere Merkmale einer Matrix zunutze, z. B. die Eigenschaft A der Matrix gemäß Definition 2.19.

Satz 2.40 *Eine Matrix $A = (a_{ij})$ mit*
(1) der Eigenschaft A,
(2) $a_{ii} \neq 0\ \forall\ i$,
lässt sich konsistent ordnen, d. h. mit einer Permutationsmatrix P ensteht eine konsistent geordnete Matrix PAP^T.

Beweis. [46]
Die Eigenschaft A der Matrix ergibt

$$PAP^T = \begin{pmatrix} \widetilde{D}_1 & M_1 \\ M_2 & \widetilde{D}_2 \end{pmatrix} = \widetilde{D}(I - \widetilde{L} - \widetilde{U}) = \widetilde{D}(I - \widetilde{J}),\ P \text{ Permutationsmatrix},$$

wobei die nicht verschwindenden Diagonalelemente von A durch die gleichzeitigen Spalten- und Zeilenvertauschungen auf der Diagonalen bleiben und die quadratischen Blöcke $\widetilde{D}_k$ bilden. Somit sind

$$\widetilde{D} = \begin{pmatrix} \widetilde{D}_1 & 0 \\ 0 & \widetilde{D}_2 \end{pmatrix} \text{ regulär und } \widetilde{L} = -\begin{pmatrix} 0 & 0 \\ \widetilde{D}_2^{-1}M_2 & 0 \end{pmatrix},\ \widetilde{U} = -\begin{pmatrix} 0 & \widetilde{D}_1^{-1}M_1 \\ 0 & 0 \end{pmatrix}.$$

Aber gerade für die Matrixform

$$\begin{pmatrix} \widetilde{D}_1 & M_1 \\ M_2 & \widetilde{D}_2 \end{pmatrix}$$

können wir zum Nachweis der konsistenten Ordnung wie im Satz 2.38 vorgehen.

$$\begin{aligned} \widetilde{J}(\alpha) &= \alpha\widetilde{L} + \alpha^{-1}\widetilde{U} \\ &= \begin{pmatrix} 0 & -\alpha^{-1}\widetilde{D}_1^{-1}M_1 \\ -\alpha\widetilde{D}_2^{-1}M_2 & 0 \end{pmatrix} \\ &= \begin{pmatrix} I_1 & 0 \\ 0 & \alpha I_2 \end{pmatrix} \begin{pmatrix} 0 & -\widetilde{D}_1^{-1}M_1 \\ -\widetilde{D}_2^{-1}M_2 & 0 \end{pmatrix} \begin{pmatrix} I_1 & 0 \\ 0 & \alpha^{-1}I_2 \end{pmatrix} \\ &= S(\alpha)\widetilde{J}(1)S(\alpha)^{-1}, \quad S(\alpha) = \mathrm{diag}(I_1, \alpha I_2). \end{aligned}$$

Damit sind $\widetilde{J}(\alpha)$ und $\widetilde{J}(1)$ ähnlich und haben die gleichen EW. Weiter sind die EW von $\widetilde{J} = \widetilde{J}(1)$ unabhängig von α. □

Die Behauptung im Satz 2.39 bleibt richtig, wenn in der Voraussetzung über die Matrix A die konsistente Ordnung durch die Eigenschaft A ersetzt wird.

Satz 2.41 *Wenn die Matrix A positiv definit ist und mit der regulären Zerlegung $A = D - C = D(I - L - U)$, $D = \mathrm{diag}(A)$, die Eigenschaft* A *besitzt, dann ist mit μ als EW von $J = D^{-1}C = L + U$ auch die Größe $-\mu$ EW von J.*

Beweis.
Es reicht nicht aus wie im Satz 2.40, die Matrix A mit ihrer Eigenschaft A auf die Form

$$B = PAP^T = \begin{pmatrix} \widetilde{D}_1 & M_1 \\ M_2 & \widetilde{D}_2 \end{pmatrix}, \quad P \text{ Permutationsmatrix},$$

zu transformieren und dann B als konsistent geordnete Matrix zu erkennen, die dann das genannte EW-Merkmal besitzt. Wir möchten und müssen die Vorzeicheneigenschaft der EW von A selber zeigen.
Dazu beschränken wir uns hier auf den Fall $\mathrm{diag}(A) = I$ und machen einige Vorbetrachtungen.

1. Mit $A > 0$ ist $a_{ii} > 0$. Eine geeignete Skalierung mit einer Diagonalmatrix der Form $A' = \hat{D}A\hat{D}$ lässt A' positiv definit und liefert $\mathrm{diag}(A') = I$.
   Die Skalierungsmatrix ist

   $$\hat{D} = \mathrm{diag}(a_{11}^{-1/2}, a_{22}^{-1/2}, ..., a_{nn}^{-1/2}).$$

Die positive Definitheit erhält man mit $x \neq 0$ aus

$$0 < x^T Ax = x^T \hat{D}^{-1} A' \hat{D}^{-1} x = x^T \hat{D}^{-T} A' \hat{D}^{-1} x = y^T A' y,\ y = \hat{D}^{-1} x \neq 0.$$

2. Die Matrizen A und A' haben an denselben Stellen Nulleinträge und NNE. Damit überträgt sich die Eigenschaft A auf die skalierte Matrix A' und es ist

$$B' = PA'P^T = \begin{pmatrix} I_1 & M_1' \\ M_2' & I_2 \end{pmatrix}, \quad \text{wobei } \dim(I_k) = \dim(\widetilde{D}_k).$$

Das lässt sich auch einfach nachrechnen gemäß

$$\begin{aligned} PA'P^T &= P\hat{D}A\hat{D}P^T = P\hat{D}P^T\, PAP^T\, P\hat{D}P^T \\ &= \begin{pmatrix} \widetilde{D}_1^{-1/2} & 0 \\ 0 & \widetilde{D}_2^{-1/2} \end{pmatrix} PAP^T \begin{pmatrix} \widetilde{D}_1^{-1/2} & 0 \\ 0 & \widetilde{D}_2^{-1/2} \end{pmatrix} \\ &= \begin{pmatrix} \widetilde{D}_1^{-1/2} & 0 \\ 0 & \widetilde{D}_2^{-1/2} \end{pmatrix} \begin{pmatrix} \widetilde{D}_1 & M_1 \\ M_2 & \widetilde{D}_2 \end{pmatrix} \begin{pmatrix} \widetilde{D}_1^{-1/2} & 0 \\ 0 & \widetilde{D}_2^{-1/2} \end{pmatrix} \\ &= \begin{pmatrix} I_1 & \widetilde{D}_1^{-1/2} M_1 \widetilde{D}_2^{-1/2} \\ \widetilde{D}_2^{-1/2} M_2 \widetilde{D}_1^{-1/2} & I_2 \end{pmatrix}. \end{aligned}$$

Sei also $D = \operatorname{diag}(A) = I$, womit $A = I - J,\ J = L + U$, ist.
Ihre Eigenschaft A bedeutet, dass die transformierte Matrix

$$B = PAP^T = PAP^{-1} = \begin{pmatrix} I_1 & M_1 \\ M_2 & I_2 \end{pmatrix}, \quad P \text{ Permutationsmatrix},$$

gebildet werden kann, die konsistent geordnet und ähnlich zu A ist.
Aber B hat auch die Darstellung

$$B = I - \widetilde{L} - \widetilde{U} = I - \widetilde{J},$$

ist konsistent geordnet, und das hat nach Satz 2.39 zur Folge, dass mit μ EW von $\widetilde{J} = \widetilde{L} + \widetilde{U}$ auch $-\mu$ EW ist.
Sei also $\sigma(\widetilde{J}) = \{\mu_1, \mu_2, ..., \mu_n\}$ das Spektrum und

$$\mu_n = -\mu_1,\ \mu_{n-1} = -\mu_2, ...,\ \mu_{[\frac{n}{2}]+1} = -\mu_{[\frac{n}{2}]}$$

(falls n ungerade ist, ist der mittlere EW $\mu_{[\frac{n+1}{2}]} = 0$).
Damit sind die EW von B gleich $1 - \mu_k$, genauso die der ähnlichen Matrix $A = I - J$. Somit hat die Matrix $J = L + U = I - A$ die EW μ_k mit der genannten Vorzeicheneigenschaft. □

In der Voraussetzung des Satzes reicht auch $D = \operatorname{diag}(A) > 0$.

Beispiel 2.27 Die Matrix

$$A = \begin{pmatrix} 1 & b & 0 \\ a & 1 & d \\ 0 & c & 1 \end{pmatrix}$$

hat die Eigenschaft A, da mit dem Permutationsvektor $p = (2, 1, 3)$ und der zugehörigen Permutationsmatrix

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

gilt

$$B = PAP^T = \left(\begin{array}{c|cc} 1 & a & d \\ \hline b & 1 & 0 \\ c & 0 & 1 \end{array}\right) = \begin{pmatrix} D_1 & M_1 \\ M_2 & D_2 \end{pmatrix} = \begin{pmatrix} I_1 & M_1 \\ M_2 & I_2 \end{pmatrix}.$$

Aus der Eigenschaft A der Matrix B folgt wie im Satz 2.41 die genannte Vorzeicheneigenschaft der EW der Ausgangsmatrix A.
Die konsistente Ordnung von A selber wird durch den Satz 2.38 gezeigt oder auf direktem Wege wie im Beispiel 2.23.

Bleibt noch die Frage zu beantworten, ob sich die Eigenschaft A einer Matrix wiederum durch einen anderen eventuell einfacheren Test nachprüfen lässt.
Eine positive Antwort lautet, dass es für irreduzible Matrizen mit nicht verschwindenden Diagonalelementen möglich ist. Der Zugang ist graphentheoretisch, wie in der Definition 2.12 für irreduzibel diagonaldominante Matrizen.
Sei $A = A(n, n) = (a_{ij}) = D(I - J),\ J = L + U$. Zur Matrix $J = (J_{ij})$ konstruiert man den Graphen $G(J)$ mit den Knoten $P_1, P_2, ..., P_n$ und den gerichteten Kanten $P_i \to P_j$, die entstehen gdw. $J_{ij} \neq 0$ ist. Da die Diagonalelemente $J_{ii} = 0$ sind, fehlen die Kanten $P_i \to P_i$.

Beispiel 2.28 Matrix, ihr Graph und Adjazenzmatrix

$$A(3,3) = \begin{pmatrix} 1 & 2 & 0 \\ -1 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}, \quad J = -\begin{pmatrix} 0 & 2 & 0 \\ -1 & 0 & 0 \\ 3 & 0 & 0 \end{pmatrix},$$

$$\text{Adjazenzmatrix } \ J_{adj} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

Graph $G(J)$: $\{P_1, P_2, P_3;\ P_1 \leftrightarrow P_2, P_3 \rightarrow P_1\}$.

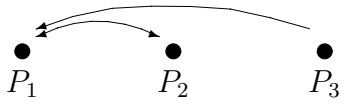


Abb. 2.2 Graph $G(J)$

Betrachten wir im Graph $G(J)$ die Längen $s_1^{(i)}, s_2^{(i)}, ...$ aller gerichteten geschlossenen Wege von P_i zu P_i, d. h. Zyklen

$$P_i \ \rightarrow\ P_{k_1} \ \rightarrow\ P_{k_2} \ \rightarrow\ ... \ \rightarrow\ P_{k_r} \ \stackrel{=}{\rightarrow}\ P_i, \quad r = s_j^{(i)}.$$

Mit dem größten gemeinsamen Teiler (GGT) bilden wir

$$l_i \ =\ GGT(s_1^{(i)}, s_2^{(i)}, ...), \quad i = 1, 2, ..., n.$$

Definition 2.27 Graph zyklisch vom Index 2
Ein Graph $G(J)$ heißt zyklisch vom Index 2, falls $l_i = 2$ für alle $i = 1, 2, ..., n$.

Definition 2.28 Matrix schwach zyklisch vom Index 2
Eine reelle Matrix A heißt schwach zyklisch vom Index 2, falls eine Permutationsmatrix P existiert, so dass

$$PAP^T \ =\ \begin{pmatrix} O_1 & M_1 \\ M_2 & O_2 \end{pmatrix}, \tag{2.42}$$

wobei $O_1(p,p),\ O_2(q,q),\ p+q=n$, Nullmatrizen sind.
Damit hat sie die Eigenschaft A *und lässt sich konsistent ordnen.*

Die genannte Eigenschaft "schwach zyklisch vom Index 2" ist somit äquivalent zur Aussage, dass die Matrix A die Eigenschaft A hat und $a_{ii} = 0$ für alle i ist.

Satz 2.42 *Eine irreduzible Matrix $A = D(I-J),\ D = \mathrm{diag}(A),\ J = L+U,\ a_{ii} \neq 0$, hat die Eigenschaft* A *gdw. der Graph $G(J)$ zyklisch vom Index 2 ist.*

Beispiel 2.29 Sei

$$A = \begin{pmatrix} 4 & -1 & 0 & -1 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ -1 & 0 & -1 & 4 \end{pmatrix}, \quad J = L + U = \frac{1}{4}\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

Der Graph ist $G(J) = \{P_1, P_2, P_3, P_4;\ P_1 \leftrightarrow P_2, P_1 \leftrightarrow P_4, P_2 \leftrightarrow P_3, P_3 \leftrightarrow P_4\}$.

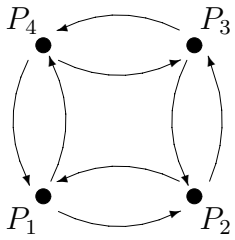


Abb. 2.3 Graph $G(J)$

Der Graph $G(J)$ ist stark zusammenhängend, d. h. für jedes Knotenpaar (P_i, P_j), $1 \le i, j \le 4$, gibt es einen gerichteten Weg $P_i \to P_{k_1} \to ... \to P_{k_m} \stackrel{=}{\to} P_j$ von P_i nach P_j. Damit ist die Matrix J irreduzibel, genauso A.

Betrachten wir von jedem Punkt aus alle Zyklen, ihre Längen und davon den GGT. Wir nehmen die wesentlichen Zyklen von $P_1 \to ... \to P_1$, also "ohne Wiederholungen" wie bei $P_1 \to P_2 \to P_3 \to P_2 \to P_3 \to P_2 \to P_1$.

Zyklus	Länge $s_j^{(1)}$
$P_1 \to P_2 \to P_1$, d. h. $P_1 \to P_{k_1} \to P_{k_2} \stackrel{=}{\to} P_1$	$s_1^{(1)} = 2$
$P_1 \to P_4 \to P_1$	$s_2^{(1)} = 2$
$P_1 \to P_2 \to P_3 \to P_2 \to P_1$	$s_3^{(1)} = 4$
$P_1 \to P_4 \to P_3 \to P_4 \to P_1$	$s_4^{(1)} = 4$
$P_1 \to P_2 \to P_3 \to P_4 \to P_3 \to P_2 \to P_1$	$s_5^{(1)} = 6$
$P_1 \to P_4 \to P_3 \to P_2 \to P_3 \to P_4 \to P_1$	$s_6^{(1)} = 6$
$P_1 \to P_2 \to P_3 \to P_4 \to P_1$	$s_7^{(1)} = 4$
$P_1 \to P_4 \to P_3 \to P_2 \to P_1$	$s_8^{(1)} = 4$

Tab. 2.1 Zyklen von $P_1 \to P_{k_1} \to ... \to P_{k_m} \stackrel{=}{\to} P_1$ mit ihren Längen $s_j^{(1)} = m$

Für $l_1 = GGT(s_j^{(1)},\ j = 1, 2, ..., 8)$ erhält man den Wert 2.
Wegen der Symmetrie im Graphen sieht man genauso leicht, dass $l_2 = l_3 = l_4 = 2$ gilt. Damit ist $G(J)$ zyklisch vom Index 2.
Folglich hat A die Eigenschaft A.

Beispiel 2.30 Sei

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad J = L + U = - \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

Der Graph ist $G(J) = \{P_1, P_2, P_3;\ P_2 \to P_1, P_3 \to P_1, P_3 \to P_2\}$.

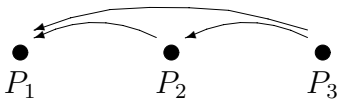


Abb. 2.4 Graph $G(J)$

Der Graph $G(J)$ ist nicht stark zusammenhängend, da es für das Knotenpaar (P_1, P_3) keinen gerichteten Weg von P_1 nach P_3 gibt. Damit ist die Matrix J reduzibel, genauso A. Es gilt

$$PAP^T = \left(\begin{array}{c|cc} 1 & 1 & 1 \\ \hline 0 & 1 & 1 \\ 0 & 0 & 1 \end{array}\right), \quad P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

A hat nicht die Eigenschaft A, sie ist aber konsistent geordnet.

Die gleichen Eigenschaften erkennen wir auch für die etwas veränderte Matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

Beispiel 2.31 Betrachten wir die Blocktridiagonalmatrix

$$A = \begin{pmatrix} D_1 & A_{12} & & & 0 \\ A_{21} & D_2 & A_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & A_{m-1,m-2} & D_{m-1} & A_{m-1,m} \\ 0 & & & A_{m,m-1} & D_m \end{pmatrix},$$

wobei alle Diagonalblöcke D_i quadratische Nullmatrizen sind.

Die Matrix $A = -L - U = -J$ ist schwach zyklisch vom Index 2, damit hat sie die Eigenschaft A. Weiterhin ist sie konsistent geordnet.

Die letzten beiden Eigenschaften sollen demonstriert werden.

Für die Eigenschaft A brauchen wir eine geeignete Permutationsmatrix. Wir wählen eine Permutation, welche die Zeilen und Spalten mit ungerader Nummer an den Anfang setzt, danach folgen der Reihe nach die Zeilen und Spalten mit gerader Nummer. Mit

$$P = \begin{pmatrix} I_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & I_3 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & I_{2k+1} \\ 0 & I_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & I_4 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & I_{2k} & 0 \end{pmatrix}, \quad m = 2k+1,$$

bzw.

$$P = \begin{pmatrix} I_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & I_3 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & I_{2k-1} & 0 \\ 0 & I_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & I_4 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & I_{2k} \end{pmatrix}, \quad m = 2k,$$

erhält man

$$PAP^T = \left(\begin{array}{cccc|cccc} D_1 & & & & A_{12} & & & \\ & D_3 & & & A_{32} & A_{34} & & \\ & & D_5 & & & A_{54} & A_{56} & \\ & & & \ddots & & & \ddots & \ddots \\ \hline A_{21} & A_{23} & & & D_2 & & & \\ & A_{43} & A_{45} & & & D_4 & & \\ & & A_{65} & \ddots & & & D_6 & \\ & & & \ddots & & & & \ddots \end{array}\right).$$

Die Transformation ist auch für beliebige Diagonalmatrizen D_i machbar.
Nach Bemerkung 2.4 ist die Matrix PAP^T konsistent geordnet. Das gilt auch für A, was wir auf anderem Weg zeigen wollen.
Es ist $A = -J = -L - U$, $Jx = \lambda x$, $x \neq 0$, $x = (u_1, u_2, ..., u_m)^T$, wobei u_i Teilvektoren entsprechender Länge gemäß der Blockstruktur von J sind.
Das EWP $Jx = \lambda x$ hat als i-te Gleichung

$$-A_{i,i-1}u_{i-1} - A_{i,i+1}u_{i+1} \;=\; \lambda u_i,\; i = 1, 2, ..., m,\; A_{10} = 0,\; A_{m,m+1} = 0,$$

woraus sich mit $\alpha \neq 0$ die Beziehungen

$$\begin{aligned} -\alpha^{i-1}A_{i,i-1}u_{i-1} - \alpha^{i-1}A_{i,i+1}u_{i+1} &= \lambda\alpha^{i-1}u_i, \\ -\alpha A_{i,i-1}(\alpha^{i-2}u_{i-1}) - \alpha^{-1}A_{i,i+1}(\alpha^i u_{i+1}) &= \lambda\alpha^{i-1}u_i \end{aligned}$$

ergeben. Somit ist $\tilde{x} = (u_1, \alpha u_2, ..., \alpha^{m-1}u_m)^T$ EV zu $J(\alpha) = \alpha L + \alpha^{-1}U$ mit dem gleichen EW λ. Aber λ ist ja unabhängig von α gewesen.

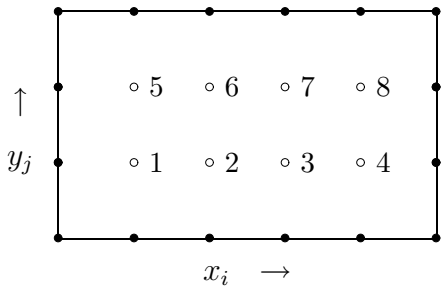
Dieses Beispiel zeigt uns, dass jede Blocktridiagonalmatrix A (2.41) mit beliebigen Diagonalmatrizen D_i die Eigenschaft A besitzt.
Was ihre konsistente Ordnung betrifft, so haben wir diese in zwei Fällen und zwar für $D_i,\ i = 1, 2, ..., m$, als
– reguläre Matrizen gemäß Satz 2.38,
– Nullmatrizen im Beispiel 2.31
nachgewiesen. In anderen Fällen sind generelle Aussagen nicht möglich.

Beispiel 2.32 Betrachten wir als letztes Beispiel die Blocktridiagonalmatrix

$$A(n,n) = \begin{pmatrix} D_1 & A_{12} & & & 0 \\ A_{21} & D_2 & A_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & A_{m-1,m-2} & D_{m-1} & A_{m-1,m} \\ 0 & & & A_{m,m-1} & D_m \end{pmatrix}, \quad n = mk,$$

wobei alle Diagonalblöcke D_i und A_{ij} quadratische $(k \times k)$-Matrizen, darüber hinaus D_i tridiagonal, $a_{ii} \neq 0\ \forall\ i$, und A_{ij} diagonal sind. Die genannten Diagonalen in den Matrizen sind mit NNE belegt.
Eine solche Matrix tritt bei der FDM wie in den Kap. 1.3, 1.4 auf.

Wir betrachten analog zum Beispiel in Kap. 1.4 ein Rechteckgebiet mit äquidistantem Gitter und $8 = 4 \cdot 2$ inneren Punkten bei zeilenweiser Nummerierung,
d. h. $n = 8,\ k = 4,\ m = 2$.



Die Unbekannten u_{ij} an den Knoten $(x_i, y_j),\ i = 1, 2, 3, 4,\ j = 1, 2$, werden sequentiell umbenannt in $u_{(j-1)k+i}$, so dass der Vektor der Unbekannten $u = (u_1, u_2, ..., u_8)^T$ ist.

Der Differenzenstern in jedem Punkt verknüpft diesen mit jeweils 4 Nachbarpunkten, wobei einige davon auf dem Rand liegen können, wo Randbedingungen vorliegen.
So erhalten wir nach einer einfachen Skalierung der Diagonale auf Eins eine Matrix mit der (2×2)-Blockstruktur

$$A = \begin{pmatrix} B & -T \\ -T & B \end{pmatrix}$$

mit (4×4)-Matrizen $B = \text{tridiag}(-\frac{1}{4}, 1, -\frac{1}{4})$ und $T = \frac{1}{4}\text{diag}(1,1,1,1)$.
A hat die Bandbreite 9.
Wir haben also das LGS $Au = b$ mit

$$A = \left(\begin{array}{cccc|cccc} a_{11} & a_{12} & & & a_{15} & & & \\ a_{21} & a_{22} & a_{23} & & & a_{26} & & \\ & a_{32} & a_{33} & a_{34} & & & a_{37} & \\ & & a_{43} & a_{44} & & & & a_{48} \\ \hline a_{51} & & & & a_{55} & a_{56} & & \\ & a_{62} & & & a_{65} & a_{66} & a_{67} & \\ & & a_{73} & & & a_{76} & a_{77} & a_{78} \\ & & & a_{84} & & & a_{87} & a_{88} \end{array}\right), \quad a_{ii} = 1, \text{ NNE } a_{ij} = -\frac{1}{4}, \ i \neq j,$$

$$J = -\left(\begin{array}{cccc|cccc} 0 & a_{12} & & & a_{15} & & & \\ a_{21} & 0 & a_{23} & & & a_{26} & & \\ & a_{32} & 0 & a_{34} & & & a_{37} & \\ & & a_{43} & 0 & & & & a_{48} \\ \hline a_{51} & & & & 0 & a_{56} & & \\ & a_{62} & & & a_{65} & 0 & a_{67} & \\ & & a_{73} & & & a_{76} & 0 & a_{78} \\ & & & a_{84} & & & a_{87} & 0 \end{array}\right)$$

und $A = I - J, \ J = L + U$.
Wir konstruieren den Graph $G(J)$.

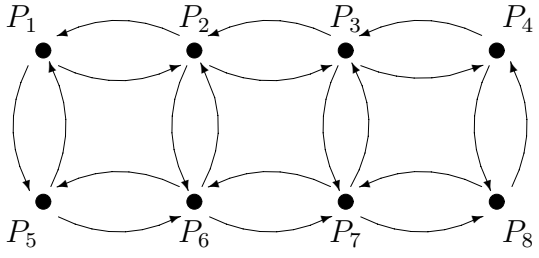


Abb. 2.5 Graph $G(J)$

Der Graph $G(A)$, der sich von $G(J)$ nur durch die zusätzlichen trivialen Zyklen $P_i \to P_i$ unterscheidet, ist stark zusammenhängend, damit ist die Matrix A irreduzibel.

Durch das paarweise Auftreten von gerichteten Verbindungen zwischen den Knoten ist $G(J)$ zyklisch vom Index 2. Somit hat die Matrix A nach Satz 2.42 die Eigenschaft A, lässt sich nach Satz 2.40 konsistent ordnen und erfüllt nach Satz 2.41 die Vorzeicheneigenschaft für ihre EW.

Man kann jedoch zeigen, dass A selbst konsistent geordnet ist. Es lässt sich leicht mit der Matrix $S = \text{diag}(1, \alpha, \alpha^2, \alpha^3,\ \alpha, \alpha^2, \alpha^3, \alpha^4)$ die Ähnlichkeitsbeziehung

$$J(\alpha) = \alpha L + \alpha^{-1} U = SJ(1)S^{-1} = SJS^{-1}, \quad \alpha \neq 0,$$

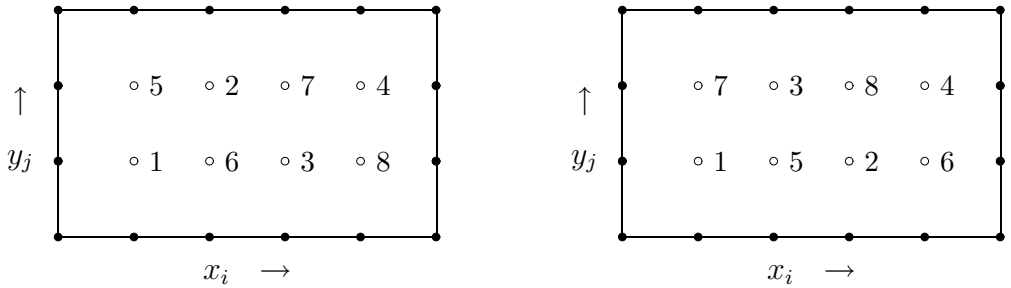
nachrechnen.

Im allgemeinen Fall eines Gitters mit k horizontalen inneren Stützstellen x_i und m vertikalen inneren Stützstellen y_j hat die Transformationsmatrix die Gestalt

$$S = \text{diag}(1, \alpha, ..., \alpha^{k-1},\ \alpha, \alpha^2, ..., \alpha^k, ...,\ \alpha^{m-1}, \alpha^m, ..., \alpha^{m-1+k-1}).$$

Es bleibt noch die Frage, ob man ohne Weiteres die Permutationsmatrix P für die Eigenschaft A angeben kann. Hier hilft uns ein kleiner Trick, der mit der Nummerierung der Knoten zusammenhängt. Man muss versuchen, in den Tridiagonalblöcken D_i die Nebendiagonalen möglichst nach "außen" zu schieben, so dass die Unbekannte u_l von der Nummerierung her 4 ferne Nachbarn erhält. Das erreicht man mit der sogenannten "Schachbrettnummerierung" (auch Schwarz-Weiß-N., Gerade-Ungerade-N.).

Zwei Varianten GN_1 bzw. GN_2 sind



Die Differenzensterne verbinden zwar dieselben Knoten, aber die Unbekannten sind anders indiziert.

Die entsprechenden Koeffizientenmatrizen des LGS haben dann folgende Strukturen mit NNE

$$A_1 = \left(\begin{array}{cccc|cccc} * & & & & * & * & & \\ & * & & & * & * & * & \\ & & * & & & * & * & * \\ & & & * & & & * & * \\ \hline * & * & & & * & & & \\ * & * & * & & & * & & \\ & * & * & * & & & * & \\ & & * & * & & & & * \end{array}\right) \text{ bzw. } A_2 = \left(\begin{array}{cccc|cccc} * & & & & * & & * & \\ & * & & & * & * & & * \\ & & * & & * & & * & * \\ & & & * & & * & & * \\ \hline * & * & * & & * & & & \\ & * & & * & & * & & \\ * & & * & & & & * & \\ & * & * & * & & & & * \end{array}\right).$$

Und schon sehen wir das Ergebnis der Eigenschaft A, nämlich

$$PAP^T = \begin{pmatrix} D_1 & M_1 \\ M_2 & D_2 \end{pmatrix}.$$

Die Umnummerierung halten wir in den Permutationsvektoren

$$p_1 = (1,6,3,8,5,2,7,4) \quad \text{bzw.} \quad p_2 = (1,5,2,6,7,3,8,4)$$

fest. Diese bilden jeweils die Permutationsmatrix P^T.
Da die Struktur von A_1 günstiger erscheint, was auch mit weniger Vertauschungen im Permutationsvektor p_1 korrespondiert, werten wir hier nur das Matrixprodukt $PAP^T,\ P = P_1$, aus.

$$PAP^T = \begin{pmatrix} 1 & & & & & & & \\ & & & 1 & & & & \\ & 1 & & & & & & \\ & & & & & & & 1 \\ & & & & 1 & & & \\ 1 & & & & & & & \\ & & & & & & 1 & \\ & & 1 & & & & & \end{pmatrix} \begin{pmatrix} * & * & & & * & & & \\ * & * & * & & & * & & \\ & * & * & * & & & * & \\ & & * & * & & & & * \\ * & & & & * & * & & \\ & * & & & * & * & * & \\ & & * & & & * & * & * \\ & & & * & & & * & * \end{pmatrix} \begin{pmatrix} 1 & & & & & & & \\ & & & & 1 & & & \\ & & 1 & & & & & \\ & & & & & & & 1 \\ & & & 1 & & & & \\ 1 & & & & & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \end{pmatrix}$$

$$= \begin{pmatrix} * & * & & * & & & & \\ & * & & * & * & * & & \\ & * & * & & * & & & \\ & & * & & & * & * & \\ * & & & * & * & & & \\ * & * & * & & * & & & \\ & & * & & & * & * & * \\ & & * & * & & & & * \end{pmatrix} \begin{pmatrix} 1 & & & & & & & \\ & & & & 1 & & & \\ & & 1 & & & & & \\ & & & & & & & 1 \\ & & & & & 1 & & \\ & 1 & & & & & & \\ & & & & & & 1 & \\ & & & 1 & & & & \end{pmatrix}$$

$$= \begin{pmatrix} * & & & * & * & & & \\ & * & & * & * & * & & \\ & & * & & * & * & * & \\ & & & * & & * & * & \\ * & * & & & * & & & \\ * & * & * & & & * & & \\ & * & * & * & & & * & \\ & & * & * & & & & * \end{pmatrix}$$

$$= A_1.$$

Die Vorgehensweise erlaubt auch die Herleitung der Matrizen und Beziehungen für allgemeine rechtwinkelige Gitter.

Wir haben hier einige Matrixeigenschaften und ihre Beziehungen zueinander untersucht. Die Ergebnisse können die Entscheidung erleichtern, mit welchen Verfahren dann die LGS zu lösen sind.

2.3 Norm

Wir betrachten reelle Vektoren $x, y, ... \in \mathbb{R}^n$ und reelle Matrizen $A, B, ... \in \mathbb{R}^{n,n}$.

2.3.1 Vektornorm

Definition 2.29 Vektornorm
Eine Vektornorm $\|x\|$ ist eine Abbildung $\mathbb{R}^n \to \mathbb{R}$ (reellwertiges Funktional) mit den folgenden drei Eigenschaften.

(1) $\|x\| \geq 0, \quad \|x\| = 0 \Leftrightarrow x = 0$ **(Positivität, Definitheit)**,

(2) $\|cx\| = |c| \ \|x\| \quad \forall c \in \mathbb{R}$ **(Homogenität)**,

(3) $\|x + y\| \leq \|x\| + \|y\|$ **(Dreiecksungleichung)**.

Die Dreiecksungleichung kann man in der Form $\|x - y\| \geq |\|x\| - \|y\||$ schreiben. Das folgt aus der Darstellung

$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|$, also $\|x\| - \|y\| \leq \|x - y\|$, sowie
$\|y\| = \|y - x + x\| \leq \|y - x\| + \|x\|$, also $-\|x - y\| \leq \|x\| - \|y\|$.

Ausgewählte Vektornormen

$\|x\|_1 = \sum_{j=1}^{n} |x_j|$ Betragssummennorm, Summennorm, Manhatten-Norm, l_1-Norm,

$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$ euklidische Norm, l_2-Norm,

$\|x\|_\infty = \max_{i=1(1)n} |x_i|$ Betragsmaximumnorm, Maximumnorm, Tschebyscheff-Norm, l_∞-Norm,

$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$ Hölder-Norm, l_p-Norm, $p \in [1, \infty)$, $p = 1, 2, \infty$ als Spezialfälle,

$\|x\|_{p,q} = \left(\sum_{i=1}^{n} q_i |x_i|^p \right)^{1/p}$ gewichtete l_p-Norm, Gewichte $q_i > 0$ gegeben,

$\|x\|_A = \sqrt{x^T A x}$ energetische Norm, Energienorm, A-Norm, wobei $A = A^T > 0, \ x^T A x = (Ax)^T x = (Ax, x) = (x, x)_A$.

Den Nachweis der drei Normeigenschaften wollen wir kurz für die Norm $\|x\|_1$ demonstrieren, wobei die Eigenschaften der Betragsfunktion genutzt werden.

(1) Es gilt natürlich $\|x\|_1 = \sum_{j=1}^{n} |x_j| \geq 0$ für alle x.

Weiter ist $\|x\|_1 = 0$ genau dann, wenn alle Beträge $|x_j|$ verschwinden und das heißt $x_j = 0 \; \forall \; j$ und damit $x = 0$.

(2) In Bezug auf die Homogenität erhalten wir

$$\|cx\|_1 = \sum_{i=1}^{n} |cx_i| = |c| \sum_{i=1}^{n} |x_i| = |c| \|x\|_1.$$

(3) Zur Dreiecksungleichung machen wir die Abschätzung

$$\|x + y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} (|x_i| + |y_i|) = \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_1 + \|y\|_1.$$

Nicht für alle Vektornormen lässt sich das so einfach zeigen.

Im Vektorraum $\mathbb{R}^n$ ist die euklidische Norm mit dem Skalarprodukt zweier Vektoren (x, y) durch die Beziehung $\|x\|_2^2 = (x, x)$ verknüpft.

Satz 2.43 Schwarzsche Ungleichung

Es gilt

$$|(x, y)| \; \leq \; \|x\|_2 \, \|y\|_2. \tag{2.43}$$

Beweis.

Wir betrachten die Ungleichung mit dem Parameter $\lambda \in \mathbb{R}$

$$0 \leq \sum_{i=1}^{n} (\lambda x_i + y_i)^2 = \lambda^2 \sum_{i=1}^{n} x_i^2 + 2\lambda \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2,$$

deren rechte Seite eine quadratische Gleichung in λ darstellt. Ihre Diskriminante erfüllt die Bedingung $\Delta \leq 0$, d. h.

$$\Delta = \Big(2 \sum_{i=1}^{n} x_i y_i \Big)^2 - 4 \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 \leq 0$$

bzw.

$$\begin{aligned} \Big(\sum_{i=1}^{n} x_i y_i \Big)^2 \; &\leq \; \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2, \\ (x, y)^2 \; &\leq \; (x, x)^2 \, (y, y)^2, \\ (x, y) \; \leq \; |(x, y)| \; &\leq \; \|x\|_2 \, \|y\|_2. \end{aligned}$$

□

Satz 2.44 *$\|x\|_1$, $\|x\|_\infty$ und $\|x\|_2$ sind Vektornormen.*

Beweis. (Skizze)
Die Teile (1) und (2) der Definition 2.29 sind für die Normen leicht zu zeigen.
Für die Norm $\|x\|_1$ haben wir Teil (3) oben schon nachgewiesen.
Also verbleiben

$$\begin{aligned}
\|x+y\|_\infty &= \max_{i=1(1)n} |x_i + y_i| \\
&\leq \max_{i=1(1)n} (|x_i| + |y_i|) \\
&= |x_{i^*}| + |y_{i^*}|, \quad 1 \leq i^* \leq n \\
&\leq \max_{i=1(1)n} |x_i| + \max_{i=1(1)n} |y_i| \\
&= \|x\|_\infty + \|y\|_\infty,
\end{aligned}$$

und

$$\begin{aligned}
\|x+y\|_2^2 &= \sum_{i=1}^{n} (x_i + y_i)^2 \\
&= \sum_{i=1}^{n} (x_i^2 + 2x_i y_i + y_i^2) \\
&= \|x\|_2^2 + \|y\|_2^2 + 2(x,y) \\
&\leq \|x\|_2^2 + \|y\|_2^2 + 2\|x\|_2\|y\|_2, \quad \text{wegen Satz 2.43} \\
&= (\|x\|_2 + \|y\|_2)^2, \\
\|x+y\|_2 &\leq \|x\|_2 + \|y\|_2.
\end{aligned}$$

□

Satz 2.45 *Die Vektornorm $\|\cdot\|$ ist eine gleichmäßig stetige Funktion auf $\mathbb{R}^n$.*

Beweis. (Skizze)
Dies ergibt sich aus der Anwendung der zweiten Darstellung der Dreiecksungleichung $|\|y\| - \|x\|| \leq \|y - x\|$ mit $y = x + h$.

(a) Stetigkeit von $\|\cdot\|$ in x
Zu zeigen ist

$$\text{in } \mathbb{R} \ : \ \forall \varepsilon > 0 \ \exists \delta > 0 \ \forall h \ \ |h| < \delta \ \rightarrow \ |f(x+h) - f(x)| < \varepsilon,$$

$$\text{in } \mathbb{R}^n : \ \forall \varepsilon > 0 \ \exists \delta > 0 \ \forall h_i \ \ |h_i| < \delta \ \rightarrow \ |f(x+h) - f(x)| < \varepsilon, \quad h = (h_1, ..., h_n)^T.$$

Mit $f(x) = \|x\|$ haben wir

$$|\|x+h\| - \|x\|| \le \|x+h-x\| = \|h\|,$$

so dass für betragskleine h_i auch $\|h\|$ klein wird und ebenso die Differenz der Funktionswerte.

(b) Bei der gleichmäßigen Stetigkeit ist zu zeigen, dass

$$\text{in } \mathbb{R}^n: \ \forall \varepsilon > 0 \ \exists \delta > 0 \ \forall h_i \ \forall x \ \ |h_i| < \delta \ \rightarrow \ |\|x+h\| - \|x\|| < \varepsilon, \quad h = (h_1, ..., h_n).$$

Es genügt

$$\delta = \frac{\varepsilon}{M}, \quad M = \sum_{i=1}^{n} \|e_i\|, \ e_i \ i\text{-ter Einheitsvektor},$$

zu nehmen, denn mit $h = (h_1, ..., h_n)^T = \sum\limits_{i=1}^{n} h_i e_i$ gilt

$$\begin{aligned} |\|x+h\| - \|x\|| \ \le \ \|h\| \ &\le \ \sum_{i=1}^{n} |h_i| \, \|e_i\| \\ &\le \ \max_{i=1(1)n} |h_i| \ \sum_{i=1}^{n} \|e_i\| \\ &= \ \frac{\varepsilon}{M} M \ = \ \varepsilon. \end{aligned}$$

□

Die Einheitssphäre der Normen

$$\mathcal{S} \ = \ \{ \ x \in \mathbb{R}^n \ : \ \|x\| = 1 \ \} \tag{2.44}$$

ist eine kompakte Menge in $\mathbb{R}^n$ (ist abgeschlossen und jeder Punkt ist ein Häufungspunkt).

Beispiel 2.33 In $\mathbb{R}^2$ und $\mathbb{R}^3$ kann man eine grafische Darstellung der Beziehung $\|x\| \le 1$ vornehmen. Für die Maximumnorm $\|x\|_\infty$ ergibt sich das Einheitsquadrat (der Einheitswürfel), für die euklidische Norm $\|x\|_2$ der Einheitskreis (die Einheitskugel) sowie für die Betragssummennorm $\|x\|_1$ ein auf der Spitze stehendes Quadrat (Oktaeder) mit der Kantenlänge $\sqrt{2}$. Daraus kann man auch die Größenverhältnisse

$$\|x\|_\infty \ \le \ \|x\|_2 \ \le \ \|x\|_1 \tag{2.45}$$

ablesen, die für beliebige Vektoren x gelten.

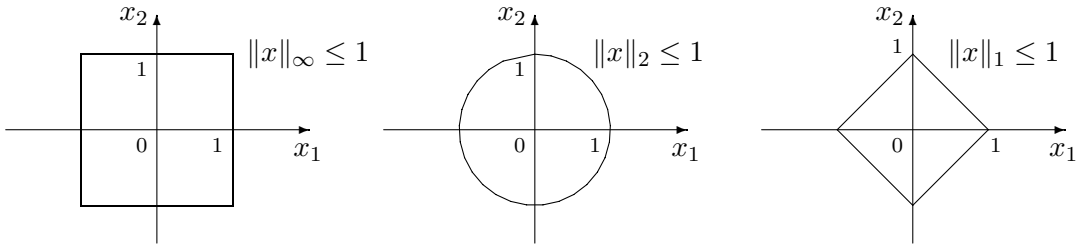


Abb. 2.6 Normverhältnisse $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$ in $\mathbb{R}^2$

Mit diesem Beispiel ist ebenfalls der Hinweis auf die **Normäquivalenz** in $\mathbb{R}^n$ verbunden.

Definition 2.30 Normäquivalenz
Für zwei beliebige Vektornormen $\|x\|_p$ und $\|x\|_q$ existieren positive Konstanten c_1 und c_2, so dass für alle Vektoren x die Ungleichung

$$c_1\|x\|_q \leq \|x\|_p \leq c_2\|x\|_q \tag{2.46}$$

gilt.

Dass bei dieser Ungleichungskette auch die Gleichheit möglich ist, lässt sich manchmal mit den Einheitsvektoren oder den einfachen Vektoren $x = (1, 1, ..., 1)^T$ überprüfen. Man sagt in diesem Fall, dass die Abschätzungen scharf sind.

Satz 2.46 *Alle Vektornormen in $\mathbb{R}^n$ sind äquivalent, d. h. für beliebige Normpaare $\|x\|_p$, $\|x\|_q$ gilt die Ungleichungskette (2.46).*

Beweis.
Für $x = 0$ ist die Beziehung (2.46) erfüllt.
Betrachten wir die Sphäre $\mathcal{S} = \{\ x \in \mathbb{R}^n\ :\ \|x\|_q = 1\ \}$.
Die stetige Funktion $f(x) = \|x\|_p$ nimmt auf dieser kompakten Menge $\mathcal{S}$ ihr Maximum und Minimum an, d. h.

$$\begin{aligned} M &= \max_{x\in\mathcal{S}} \|x\|_p, \\ m &= \min_{x\in\mathcal{S}} \|x\|_p. \end{aligned}$$

Sei $x \neq 0$. Dann ist $x/\|x\|_q \in \mathcal{S}$ und man erhält unter Verwendung der Homogenität der Norm

$$m \leq \left\| \frac{x}{\|x\|_q} \right\|_p = \frac{\|x\|_p}{\|x\|_q} \leq M,$$

also

$$m\|x\|_q \leq \|x\|_p \leq M\|x\|_q, \quad \text{mit } m = c_1,\ M = c_2.$$

□

Satz 2.47 Normäquivalenzen

Es gelten in $\mathbb{R}^n$ die Normäquivalenzen

(1) $\|x\|_\infty \le \|x\|_1 \le n\,\|x\|_\infty$,

(2) $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\,\|x\|_\infty$,

(3) $\frac{1}{n}\,\|x\|_1 \le \|x\|_\infty \le \|x\|_1$,

(4) $\frac{1}{\sqrt{n}}\,\|x\|_2 \le \|x\|_\infty \le \|x\|_2$,

(5) $\frac{1}{n}\,\|x\|_1 \le \|x\|_2 \le \sqrt{n}\,\|x\|_1$,

(6) $\frac{1}{\sqrt{n}}\,\|x\|_2 \le \|x\|_1 \le n\,\|x\|_2$,

(7) $\frac{1}{\sqrt{n}}\,\|x\|_1 \le \|x\|_2 \le \|x\|_1$ *(besser als (5))*,

(8) $\|x\|_2 \le \|x\|_1 \le \sqrt{n}\,\|x\|_2$ *(besser als (6))*.

Beweis. (Skizze)
Viele Beziehungen lassen sich mittels der Betragseigenschaften, Binomialformel und Schwarzschen Ungleichung (2.43) nachweisen.
Wir demonstrieren dies für die Ungleichungskette (8).

Es gilt $\|x\|_1 \le \sqrt{n}\,\|x\|_2$, weil

$$
\begin{aligned}
0 &\le \sum_{\substack{i,j=1\\ i<j}}^{n} (|x_i| - |x_j|)^2,\\
0 &\le (n-1)\sum_{i=1}^{n} x_i^2 - 2\sum_{\substack{i,j=1\\ i<j}}^{n} |x_i x_j|,\\
0 &\le n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i^2 + 2\sum_{\substack{i,j=1\\ i<j}}^{n} |x_i x_j|\Big),\\
0 &\le n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} |x_i|\Big)^2,\\
\Big(\sum_{i=1}^{n} |x_i|\Big)^2 &\le n\sum_{i=1}^{n} x_i^2, \quad \text{d. h.} \quad \sum_{i=1}^{n} |x_i| \le \sqrt{n}\,\sqrt{\sum_{i=1}^{n} x_i^2}.
\end{aligned}
$$

Es gilt $\|x\|_2 \le \|x\|_1$, weil

$$
\sum_{i=1}^{n} x_i^2 \le \sum_{i=1}^{n} x_i^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{n} |x_i|\,|x_j| = \Big(\sum_{i=1}^{n} |x_i|\Big)^2, \quad \text{d. h.} \quad \sqrt{\sum_{i=1}^{n} x_i^2} \le \sum_{i=1}^{n} |x_i|.
$$

□

Wir notieren einige Normen für ausgewählte Vektoren.

$\|\cdot\|$	$x=(1,0,...,0)^T$	$x=(1,1,...,1)^T$
$\|x\|_\infty$	1	1
$\|x\|_2$	1	$\sqrt{n}$
$\|x\|_1$	1	n

Tab. 2.2 Normen für ersten Einheitsvektor und Einsvektor

Damit werden aus den Ungeichungsketten

$$\|x\|_\infty \le \|x\|_2 \le \|x\|_1 \quad \text{und} \quad \frac{1}{\sqrt{n}}\|x\|_1 \le \|x\|_2 \le \sqrt{n}\,\|x\|_\infty$$

überall Gleichheiten.
Wir geben einige grafische Darstellungen von ausgewählten Normäquivalenzen in $\mathbb{R}^2$ an.

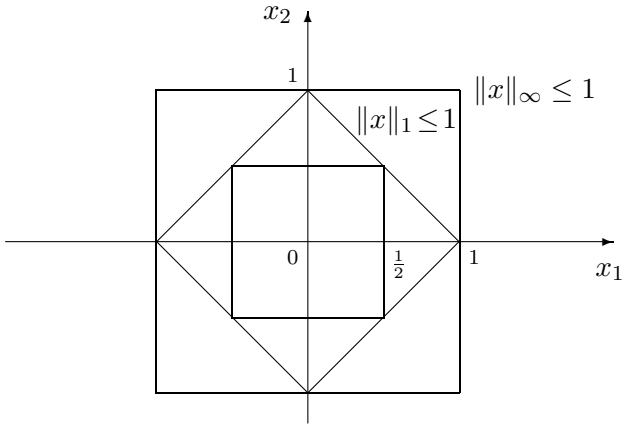


Abb. 2.7 Normäquivalenz $\|x\|_\infty \le \|x\|_1 \le 2\,\|x\|_\infty$ in $\mathbb{R}^2$

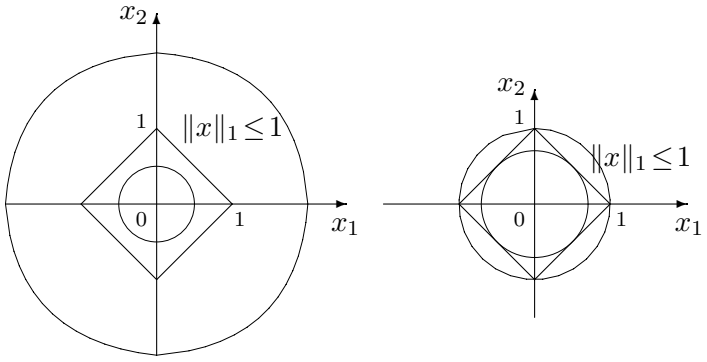


Abb. 2.8 $\frac{1}{\sqrt{2}}\|x\|_2 \le \|x\|_1 \le 2\,\|x\|_2$, $\|x\|_2 \le \|x\|_1 \le \sqrt{2}\,\|x\|_2$ in $\mathbb{R}^2$,
grobe Abschätzung, feine Abschätzung

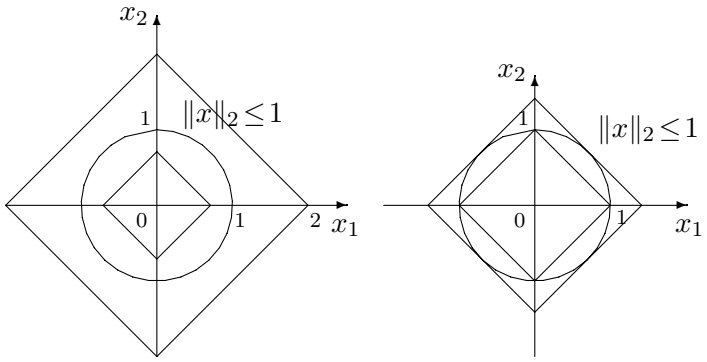


Abb. 2.9 $\frac{1}{2}\|x\|_1 \leq \|x\|_2 \leq \sqrt{2}\,\|x\|_1$ (grob), $\frac{1}{\sqrt{2}}\|x\|_1 \leq \|x\|_2 \leq \|x\|_1$ (fein) in $\mathbb{R}^2$

2.3.2 Matrixnorm

Im Vektorraum $\mathbb{R}^n$ sei eine Norm $\|\cdot\|$ gegeben.
Mit der Matrix $A \in \mathbb{R}^{n,n}$ wird eine eindeutige Abbildung von $\mathbb{R}^n$ in $\mathbb{R}^n$ vermittelt. Sie ist linear, denn es gilt

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay \quad \text{für alle } x, y \in \mathbb{R}^n,\ \alpha, \beta \in \mathbb{R}. \tag{2.47}$$

Die Abbildung A heißt beschränkt, falls eine Konstante $C > 0$ existiert, so dass für alle $x \in \mathbb{R}^n$ gilt

$$\|Ax\| \ \leq\ C\,\|x\|. \tag{2.48}$$

Definition 2.31 Induzierte Norm einer quadratischen Matrix
Sei $\|x\|$ eine Vektornorm.
Diese induziert die Matrixnorm $\|A\|$ (zugeordnete, natürliche Norm, Grenzennorm, zugehörige Operatornorm)

$$\|A\| \ =\ \inf\{\ C\ : \|Ax\| \ \leq\ C\|x\| \ \ \forall\ x \neq 0\ \}. \tag{2.49}$$

Das Infimum der Menge aller Schranken in (2.48), also die kleinste derartige Schranke, ist die induzierte Matrixnorm, eine charakteristische Zahl für die Matrix. Für diese gilt

$$\|Ax\| \leq \|A\|\,\|x\| \ \ \forall\ x. \tag{2.50}$$

Die Vektornorm heißt Induzierende. Die induzierte Matrixnorm ist mit der zu Grunde liegenden Vektornorm zugleich **kompatibel** (siehe Definition 2.32) und unter allen mit dieser Vektornorm kompatiblen Matrixnormen die kleinste.
Die anschauliche Bedeutung der induzierten Matrixnorm ist die maximale Streckung, die ein Vektor x durch die Abbildung A erfährt.
Die induzierte Matrixnorm ist wohl definiert im Sinne einer Norm, was leicht nachzuprüfen ist, wenn man die Eigenschaften der Vektornorm ausnutzt.

Definition 2.32 Kompatible Matrixnorm, Kompatibilitätsbedingung
Eine Matrixnorm $\|A\|_M$ heißt kompatibel (passend, verträglich, konsistent) zu einer Vektornorm $\|x\|_V$, wenn für alle A und x gilt

$$\|Ax\|_V \;\leq\; \|A\|_M \, \|x\|_V. \tag{2.51}$$

Ist die Abschätzung scharf, d. h., es gilt das Gleichheitszeichen für irgendeinen Nichtnullvektor, dann handelt es sich bei der kompatiblen Norm um die induzierte.

In der Literatur findet man auch die folgende äquivalente Definition der induzierten Matrixnorm.

Satz 2.48 *Für die induzierte Matrixnorm gilt*

$$\|A\| \;=\; \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \;=\; \max_{\|x\|=1} \|Ax\|. \tag{2.52}$$

Beweis.
Wir zeigen zuerst die linke Gleichheit.
Sei $s = \sup\limits_{x\neq 0}(\|Ax\|/\|x\|)$. Offenbar ist dann $\|Ax\|/\|x\| \leq s$ für alle $x \neq 0$,
also $\|Ax\| \leq s\|x\|$ für alle $x \in \mathbb{R}^n$. Nach Definition 2.31 ist damit $\|A\| \leq s$.

Aus der Supremumeigenschaft folgt, dass für beliebiges $\varepsilon > 0$ ein Element $z \neq 0$ existiert, so dass $s - \varepsilon \leq \|Az\|/\|z\|$ gilt bzw.

$$(s-\varepsilon)\|z\| \leq \|Az\| \leq C\,\|z\|.$$

Folglich ist $s - \varepsilon \leq C$ für alle Schranken C, also $s - \varepsilon \leq \inf C = \|A\|$.
Da $\varepsilon > 0$ beliebig war, ist $s \leq \|A\|$.
Aus beiden Ungleichungen ergibt sich die Gleichheit.
Mit den Normeigenschaften erhält man schließlich

$$\sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x\neq 0} \Big\|A\Big(\frac{x}{\|x\|}\Big)\Big\| = \sup_{\|y\|=1} \|Ay\| = \max_{\|y\|=1} \|Ay\|.$$

unter Verwendung der Kompaktheit der Menge $\{y : \|y\| = 1\}$ und der Stetigkeit der Norm. □

Eigenschaften von Matrixnormen

(a) $\|A\| \geq 0, \quad \|A\| = 0 \Leftrightarrow A = 0$ (**Positivität, Definitheit**),

(b) $\|cA\| = |c|\,\|A\| \quad \forall c \in \mathbb{R}$ (**Homogenität**),

(c) $\|A + B\| \leq \|A\| + \|B\|$ (**Dreiecksungleichung**),

(d) $\|AB\| \leq \|A\|\,\|B\|$ (**Multiplikativität, Submultiplikativität**),

(e) $\|Ax\| \leq \|A\|\,\|x\|$ (mit kompatiblen Normen).

Aus der Eigenschaft (d) folgt insbesondere $\|A^k\| \leq \|A\|^k$ für beliebiges $k \in \mathbb{N}$.

Ausgewählte Matrixnormen für $A = (a_{ij})_{i,j=1}^{n}$

$$\begin{aligned}
\|A\|_\infty \;=\; \|A\|_Z \quad &= \quad \max_{i=1(1)n} \sum_{j=1}^{n} |a_{ij}| && \text{Zeilensummennorm,}\\
\|A\|_1 \;=\; \|A\|_S \quad &= \quad \max_{j=1(1)n} \sum_{i=1}^{n} |a_{ij}| && \text{Spaltensummennorm,}\\
\|A\|_F \;=\; \|A\|_E \quad &= \quad \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2} && \text{Frobenius-Norm, Schursche Norm,}\\
&= \quad \sqrt{\mathrm{spur}(A^T A)} && \text{euklidische Norm,}\\
\|A\|_G \;=\; \|A\|_{max} \quad &= \quad n \max_{i,j=1(1)n} |a_{ij}| && \text{Gesamtnorm, Maximumnorm,}\\
\|A\|_2 \quad &= \quad \sqrt{\max_{i=1(1)n} \mu_i}\,, && 0 \le \mu_i \in \sigma(A^T A) \text{ (Spektrum),}\\
&= \quad \sqrt{\rho(A^T A)} && \text{Spektralnorm, Hilbert-Norm,}\\
& && \text{größter Singulärwert von } A.
\end{aligned}$$

Wegen $\sigma(A^T A) = \sigma(AA^T)$ (siehe Kap. 2.2.2 Punkt 4) kann man als Spektralnorm auch $\|A\|_2 = \sqrt{\rho(AA^T)}$ nehmen.
Falls $A = A^T$ ist (bzw. hermitesche Matrix im komplexen Fall), dann gilt

$$\|A\|_2 = \sqrt{\rho(A^2)} = \rho(A) = \max_{i=1(1)n} |\lambda_i(A)|. \tag{2.53}$$

Für beliebige Matrizen stellt der Spektralradius $\rho(A)$ eine untere Schranke aller Matrixnormen dar (Abschätzung (2.55)).
Frobenius- und Spektralnorm sind invariant unter der Orthogonaltransformation. Mit einer orthogonalen Matrix Q $(Q^T Q = I)$ gilt nämlich

$$(QA)^T(QA) \;=\; A^T(Q^T Q)A \;=\; A^T A.$$

Vektornorm	Matrixnorm	
$\|x\|$	kompatible	induzierte
$\|x\|_1$	$\|A\|_1,\ \|A\|_G$	$\|A\|_1$
$\|x\|_2$	$\|A\|_2,\ \|A\|_G,\ \|A\|_F$	$\|A\|_2$
$\|x\|_\infty$	$\|A\|_\infty,\ \|A\|_G$	$\|A\|_\infty$

Tab. 2.3 Kompatible und induzierte Matrixnormen

Die Frobenius-Norm stellt keine induzierte Matrixnorm dar.

Wir wollen beispielhaft zeigen, dass die Zeilensummennorm durch die Maximumnorm des Vektors und die Spaltensummennorm durch die Summennorm des Vektors induziert werden, sowie einige Aussagen über die euklidische Norm im Zusammenhang mit Spektral- und Frobenius-Norm machen.

(a) Für A als Nullmatrix gilt natürlich die Beziehung $\|Ax\|_\infty \le \|A\|_\infty \|x\|_\infty$.
Als nächstes zeigt man die Kompatibilität gemäß

$$\begin{aligned}\|Ax\|_\infty &= \max_{i=1(1)n} \Big| \sum_{j=1}^{n} a_{ij}x_j \Big| \le \max_{i=1(1)n} \sum_{j=1}^{n} |a_{ij}x_j| \le \max_{i=1(1)n} \sum_{j=1}^{n} \Big(|a_{ij}| \max_{k=1(1)n} |x_k| \Big) \le \\ &\le \Big(\max_{i=1(1)n} \sum_{j=1}^{n} |a_{ij}| \Big) \max_{k=1(1)n} |x_k| = \|A\|_\infty \|x\|_\infty .\end{aligned}$$

Der Fall der Gleichheit wird mit einem speziell definierten Vektor x überprüft.
Sei $\|A\|_\infty = \sum_{j=1}^{n} |a_{kj}| = \max_{i=1(1)n} \sum_{j=1}^{n} |a_{ij}|$ und

$$x = (x_1, x_2, ..., x_n)^T \text{ mit } x_j = \begin{cases} 1, & \text{falls } a_{kj} = 0, \\ |a_{kj}|/a_{kj}, & \text{falls } a_{kj} \neq 0. \end{cases}$$

Dann gelten für diesen Vektor $\|x\|_\infty = 1$,

$$\Big| \sum_{j=1}^{n} a_{kj}x_j \Big| = \Big| \sum_{j=1}^{n} a_{kj} \frac{|a_{kj}|}{a_{kj}} \Big| = 1 \cdot \sum_{j=1}^{n} |a_{kj}| = \|A\|_\infty \|x\|_\infty \text{ und } \|Ax\|_\infty = \|A\|_\infty \|x\|_\infty .$$

(b) Für A als Nullmatrix gilt natürlich die Beziehung $\|Ax\|_1 \le \|A\|_1 \|x\|_1$.
Als nächstes zeigt man die Kompatibilität gemäß

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^{n} \Big| \sum_{j=1}^{n} a_{ij}x_j \Big| \le \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| \, |x_j| = \sum_{j=1}^{n} \Big(|x_j| \sum_{i=1}^{n} |a_{ij}| \Big) \le \\ &\le \sum_{j=1}^{n} \Big(|x_j| \max_{j=1(1)n} \sum_{i=1}^{n} |a_{ij}| \Big) = \Big(\max_{j=1(1)n} \sum_{i=1}^{n} |a_{ij}| \Big) \sum_{j=1}^{n} |x_j| = \|A\|_1 \|x\|_1 .\end{aligned}$$

Der Fall der Gleichheit wird mit einem speziell definierten Vektor x überprüft.
Sei $\|A\|_1 = \sum_{i=1}^{n} |a_{ik}| = \max_{j=1(1)n} \sum_{i=1}^{n} |a_{ij}|$ und

$$x = (x_1, x_2, ..., x_n)^T \text{ mit } x_j = \begin{cases} 1, & \text{falls } j = k, \\ 0, & \text{sonst.} \end{cases}$$

Dann gelten für diesen Vektor $\|x\|_1 = 1$ und

$$\|Ax\|_1 = \sum_{i=1}^{n} \Big| \sum_{j=1}^{n} a_{ij}x_j \Big| = \sum_{i=1}^{n} |a_{ik}| = \|A\|_1 = \|A\|_1 \cdot 1 = \|A\|_1 \|x\|_1 .$$

(c) Die Spektralnorm $\|A\|_2$ wird induziert durch $\|x\|_2$.
Es gelten $B = A^TA = B^T \geq 0$, $\mu_i(B) \geq 0$, $\mu_1 \leq \mu_2 \leq ... \leq \mu_n$ und

$$\|A\|_2 = \sqrt{\max_{i=1(1)n} \mu_i} = \sqrt{\mu_n}.$$

Nach Satz 2.31 ist B diagonalisierbar und hat n zugehörige orthonormierte EV v_i, $i = 1, 2, ..., n$, die eine Basis von $\mathbb{R}^n$ sind.
Für $x = 0$ ist die Ungleichung (2.51) erfüllt. Sei

$$x = \sum_{i=1}^{n} c_i v_i, \quad \text{mit } \|x\|_2 = \sqrt{(x,x)} = \sqrt{\sum_{i=1}^{n} c_i^2} = 1.$$

Daraus folgen

$$\begin{aligned}
\|Ax\|_2^2 &= (Ax, Ax) = (Ax)^T(Ax) = x^TBx \\
&= \Big(\sum_{i=1}^{n} c_i v_i^T\Big)\Big(\sum_{i=1}^{n} c_i B v_i\Big) \\
&= \sum_{i,j=1}^{n} c_i c_j \mu_j v_i^T v_j \\
&= \sum_{i=1}^{n} \mu_i c_i^2 \\
&\leq \mu_n \sum_{i=1}^{n} c_i^2 \\
&\leq \mu_n \|x\|_2^2, \\
\|Ax\|_2 &\leq \sqrt{\mu_n}\|x\|_2 = \|A\|_2\|x\|_2 \quad \text{für } \|x\|_2 = 1.
\end{aligned}$$

Für beliebiges $x \neq 0$ macht man zunächst die Normierung $x/\|x\|_2$, um dann aus

$$\Big\|A\Big(\frac{x}{\|x\|_2}\Big)\Big\|_2 \leq \sqrt{\mu_n}$$

das gewünschte Ergebnis zu erhalten.
Die Gleichheit kann man mit $v_n \neq 0$ nachweisen.

$$\|Av_n\|_2^2 = v_n^TBv_n = v_n^T\mu_n v_n = \mu_n \cdot 1 = \mu_n\|v_n\|_2^2.$$

(d) Die Frobenius-Norm $\|A\|_F$ ist kompatibel mit $\|x\|_2$.
Unter Verwendung der Schwarzschen Ungleichung (2.43) folgt

$$\begin{aligned}\|Ax\|_2^2 &= \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} a_{ij}x_j\Big)^2 = \sum_{i=1}^{n}(a_i,x)^2 \le \sum_{i=1}^{n}\|a_i\|_2^2\|x\|_2^2 = \\ &= \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2\Big)\|x\|_2^2 = \|A\|_F^2\|x\|_2^2.\end{aligned}$$

(e) Die Gesamtnorm $\|A\|_G$ ist kompatibel mit $\|x\|_1$.

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^{n}\Big|\sum_{j=1}^{n} a_{ij}x_j\Big| \le \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|\,|x_j| = \sum_{j=1}^{n}\Big(|x_j|\sum_{i=1}^{n}|a_{ij}|\Big) \le \\ &\le \sum_{j=1}^{n}\Big(|x_j|\; n\max_{i,j=1(1)n}|a_{ij}|\Big) = n\max_{i,j=1(1)n}|a_{ij}|\sum_{j=1}^{n}|x_j| = \|A\|_G\|x\|_1.\end{aligned}$$

Sofortiges Ergebnis der Definition 2.31 ist der folgende Satz.

Satz 2.49
Für die Einheitsmatrix I gilt bei jeder beliebigen induzierten Matrixnorm $\|I\| = 1$.

Damit ist leicht die Frage zu beanworten, welche Matrixnormen also nicht induziert sein können, nämlich z. B. $\|A\|_G$ und $\|A\|_F$.

Werden Matrixnormen benötigt, wie bei der Untersuchung des Konvergenzverhaltens von IV zur Lösung von LGS, sollte man, wenn es ausreicht, auf die einfach zu berechnenden Zeilen- oder Spaltensummennorm zurückgreifen.

Beispiel 2.34
$$A = \begin{pmatrix} 1 & -3 \\ -5 & 2 \end{pmatrix}.$$

- Zeilen- bzw. Spaltensummennorm

$$\begin{aligned}\|A\|_\infty &= \max(|1|+|-3|, |-5|+|2|) = 7, \\ \|A\|_1 &= \max(|1|+|-5|, |-3|+|2|) = 6.\end{aligned}$$

- Frobenius-Norm

$$\|A\|_F = \sqrt{1^2+3^2+5^2+2^2} = \sqrt{39} = 6.244\,998\ldots\,.$$

- Gesamtnorm

$$\|A\|_G = 2\max(|1|,|-3|,|-5|,|2|) = 10.$$

- Spektralnorm

$$A^TA = \begin{pmatrix} 26 & -13 \\ -13 & 13 \end{pmatrix}, \quad 0 = \det(A^TA - \lambda I) = p_2(\lambda) = \lambda^2 - 39\lambda + 169.$$

Daraus folgen

$$\begin{aligned} \lambda_{1,2} &= \frac{13}{2}(3 \pm \sqrt{5}) > 0, \quad \text{als Singulärwerte } \sigma_i = \sqrt{\lambda_i} \text{ von } A \text{ und} \\ \|A\|_2 &= \sqrt{\max(\lambda_1, \lambda_2)} = 5.833\,904... \ . \end{aligned}$$

Der Wert $\max\limits_{i,j=1(1)n} |a_{ij}|$ ist keine Norm, weil die Submultiplikativität nicht erfüllt ist.
Für die Matrizen

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \ B = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \text{ und } AB = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

wäre $\|AB\| = 2$ und $\|A\| \, \|B\| = 1$ und damit $\|AB\| \not\leq \|A\| \, \|B\|$.

Beziehungen zwischen den Matrixnormen

Unter Beachtung von Satz 2.34 erhält man

$$\|A\|_2^2 = \max_{i=1(1)n} \mu_i(A^TA) \leq \sum_{i=1}^{n} \mu_i(A^TA) = \text{spur}(A^TA) = \|A\|_F^2.$$

und somit $\|A\|_2 \leq \|A\|_F$.
Andererseits ist auch

$$n \, \|A\|_2^2 = n \max_{i=1(1)n} \mu_i(A^TA) \geq \sum_{i=1}^{n} \mu_i(A^TA) = \|A\|_F^2,$$

was die Ungleichung $\|A\|_F \leq \sqrt{n} \, \|A\|_2$ ergibt.
Die Ungleichungskette lautet damit

$$\frac{1}{\sqrt{n}} \, \|A\|_F \leq \|A\|_2 \leq \|A\|_F \leq \sqrt{n} \, \|A\|_2.$$

Auf ähnliche Weise erhält man auch für die anderen Matrixnormen solche Beziehungen.

Satz 2.50 Äquivalenz von Matrixnormen

Für jedes Paar von Matrixnormen $\|A\|_p, \|A\|_q$ existieren positive Konstanten c_1, c_2, so dass für alle Matrizen A die Ungleichung

$$c_1\|A\|_p \ \leq \ \|A\|_q \ \leq \ c_2\|A\|_p \tag{2.54}$$

gilt.

2.3.3 EWP und Norm einer Matrix

Eine beliebige Matrixnorm stellt eine obere Schranke für die Beträge der EW dar. Wenn wir das EWP $Ax = \lambda x,\ x \neq 0$, und kompatible Normen haben, dann gilt

$$|\lambda|\,\|x\| = \|\lambda x\| = \|Ax\| \leq \|A\|\,\|x\| \quad \text{und somit} \quad |\lambda| \leq \|A\|. \tag{2.55}$$

Für den Spektralradius bedeutet dies $\rho(A) \leq \|A\|$.
Wenn die EW der Matrix betragsgroß sind, dann sind es auch die Matrixnormen. Ein großer Wert $\|A\|_F$ heißt, dass die Matrix A auch betragsgroße Elemente hat.

Satz 2.51 Satz von Cayley-Hamilton

Wenn $p_n(\lambda) = \det(A - \lambda I) = 0$ die charakteristische Gleichung der Matrix A ist, dann gilt $p_n(A) = 0$ (Nullmatrix).

Beweis. Wir machen den Beweis nur in zwei Sonderfällen.
Das charakteristische Polynom zu A hat die Normalform

$$p_n(\lambda) = c_0\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + ... + c_n, \quad c_0 = (-1)^n.$$

Wegen λ^k EW von A^k hat $p_n(A)$ den EW $\sum_{i=0}^{n} c_i\lambda^{n-i} = p_n(\lambda) = 0$.

(1) Sei $A = A^T$.
Damit ist $p_n(A) = p_n(A)^T$.
Seien λ_i die EW von A mit den zugehörigen linear unabhängigen EV v_i, die auch EV von $p_n(A)$ sind.
Es gelten

$$\begin{aligned}
p_n(A)v_i &= p_n(\lambda_i)v_i = 0 \cdot v_i = 0,\ i = 1, 2, ..., n, \\
\|p_n(A)\|_2^2 &= \max_{i=1(1)n} \mu_i(p_n(A)^T p_n(A)), \quad \mu_i \text{ EW} \\
&= \rho(p_n(A)^2) \\
&= [\rho(p_n(A))]^2 \\
&= 0
\end{aligned}$$

und mit der Definitheit der Norm $p_n(A) = 0$.

(2) A hat n linear unabhängige EV.
Wir haben

$$Av_i = \lambda_i v_i, \quad i = 1, 2, ..., n,$$

mit $\{v_i\}$ als Basis des Raums.

Daraus folgen

$$\begin{aligned} p_n(A)v_i &= p_n(\lambda_i)v_i = 0 \cdot v_i = 0, \quad i = 1, 2, ..., n, \\ \sum_{i=0}^{n} p_n(A)c_i v_i &= 0, \quad c_i \in \mathbb{R}, \\ p_n(A)\sum_{i=0}^{n} c_i v_i &= 0, \\ p_n(A)x &= 0 \text{ für alle } x \in \mathbb{R}^n, \end{aligned}$$

was nur für $p_n(A) = 0$ erfüllbar ist. □

Für Konvergenzuntersuchungen von IV sind Abschätzungen der Norm der Iterationsmatrix oder ihres Spektralradius wichtig. Deshalb ist folgender Satz sehr nützlich.

Satz 2.52 Matrixnorm und Spektralradius

Sei $\rho(A)$ der Spektralradius der Matrix $A = A(n, n)$.

(1) Für jedes $\varepsilon > 0$ gibt es eine Matrixnorm $\|\cdot\|$ mit

$$\|A\| \leq \rho(A) + \varepsilon. \tag{2.56}$$

Insbesondere folgt daraus: ist $\rho(A) < 1$, dann gilt $\|A\| < 1$.

(2) Wenn jeder EW $\lambda(A)$, für den $|\lambda| = \rho(A)$ nur lineare Elementarteiler besitzt, d. h., seine Vielfachheit stimmt mit der Anzahl der zugehörigen linear unabhängigen EV überein, dann existiert eine Norm mit $\|A\| = \rho(A)$.

(3) Es gilt für jede Matrix A mit beliebiger Matrixnorm

$$\rho(A) = \lim_{m\to\infty} \sqrt[m]{\|A^m\|}. \tag{2.57}$$

Beweis. [18]

(1) Nach Satz 2.32 ist die Matrix $A(n, n)$ ähnlich zu einer Matrix in Jordan-Normalform $J = T^{-1}AT$. $J = \text{diag}(J_1, J_2, ..., J_p)$ ist eine Blockdiagonalmatrix mit den quadratischen t-dimensionalen Jordan-Zellen und den EW von A auf der Diagonalen

$$J_j(t, t)(\lambda) = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda \end{pmatrix}, \quad j = 1, 2, ..., p \leq n.$$

Sei $D_\varepsilon = \text{diag}(1, \varepsilon, \varepsilon^2, ..., \varepsilon^{n-1}),\ \varepsilon > 0$ beliebig.
Dann hat die zu J ähnliche Matrix $\widetilde{J} = D_\varepsilon^{-1} J D_\varepsilon$ die Jordan-Zellen

$$\widetilde{J}_j(t,t)(\lambda) = \begin{pmatrix} \lambda & \varepsilon & & & \\ & \lambda & \varepsilon & & \\ & & \ddots & \ddots & \\ & & & \ddots & \varepsilon \\ & & & & \lambda \end{pmatrix}, \quad j = 1, 2, ..., p \leq n.$$

Die Zeilensummennorm für $\widetilde{J}$ erfüllt

$$\rho(A) + \varepsilon = \max|\lambda(A)| + \varepsilon \geq \|\widetilde{J}\|_\infty = \|D_\varepsilon^{-1} J D_\varepsilon\|_\infty = \|(TD_\varepsilon)^{-1} A T D_\varepsilon\|_\infty = \|S^{-1}AS\|_\infty,$$

wobei $S = TD_\varepsilon$ regulär ist.
Die gesuchte Norm für die Matrix A ist also $\|A\|_s = \|S^{-1}AS\|_\infty$, die auch der Ungleichung $\|A\|_s \leq \rho(A) + \varepsilon$ genügt.
Wir finden noch die zugehörige (induzierende) Vektornorm $\|\cdot\|_v$ aus der Bedingung $\|Ax\|_v \leq \|S^{-1}AS\|_\infty \|x\|_v$. Dazu bemerkt man, dass mit $x = Sy$ und Tabelle 2.3 die Beziehung

$$\|S^{-1}Ax\|_\infty = \|S^{-1}ASy\|_\infty \leq \|S^{-1}AS\|_\infty \|y\|_\infty = \|S^{-1}AS\|_\infty \|S^{-1}x\|_\infty$$

gilt und somit sich die Vektornorm $\|\cdot\|_v = \|S^{-1}\cdot\|_\infty$ ergibt.

(2) Seien $\lambda_i,\ i = 1, 2, ..., n$, die EW von A und

$$\rho(A) = \max|\lambda_i| = |\lambda_1| = |\lambda_2| = ... = |\lambda_k| > |\lambda_{k+1}| \geq ... \geq |\lambda_n|.$$

Lineare Elementarteiler zu den EW $\lambda_i,\ i = 1, 2, ..., k$, bedeuten, dass die Jordan-Zellen $J_i(t,t)(\lambda_i)$ die Dimension $t = 1$ haben.
Wir wählen nun $\varepsilon = \rho(A) - |\lambda_{k+1}| > 0$.
Dann gilt

$$\begin{aligned} \rho(A) \leq \|A\|_s &= \|(TD_\varepsilon)^{-1} A T D_\varepsilon\|_\infty \\ &\leq \max(\rho(A), |\lambda_{k+2}| + \rho(A) - |\lambda_{k+1}|, ..., |\lambda_n| + \rho(A) - |\lambda_{k+1}|) \\ &= \rho(A), \end{aligned}$$

also $\|A\|_s = \|S^{-1}AS\|_\infty = \rho(A)$.
Im Sonderfall $k = n$ haben alle EW nur lineare Elementarteiler und die Matrix A wird auf eine ähnliche Diagonalform $J = T^{-1}AT$ transformiert, so dass ohne Verwendung von D_ε sofort $\|A\|_s = \|T^{-1}AT\|_\infty = \rho(A)$ gilt.

(3) Eine kurze Plausibilitätsbetrachtung dazu lautet:
Der Grenzwert ergibt sich wegen $|\lambda(A)| \leq \|A\|$ und der Dominanz des betragsgrößten EW gegenüber den anderen bei der Bildung der Potenzen A^m mit wachsendem m.

Wir gehen zum Nachweis des Grenzwertes (2.57) schrittweise vor.
Bekannt sind die folgenden Beziehungen und Aussagen.

$$\begin{aligned}
\rho(A) &= \max\{|\lambda| \ : \ Ax = \lambda x, \ x \neq 0\}, \\
|\lambda| &\leq \rho(A) \leq \|A\|, \ \ \|A^m\| \leq \|A\|^m, \\
\rho(A^m) &= [\rho(A)]^m, \ \ \rho(A) = \sqrt[m]{\rho(A^m)}, \\
\|A\| &\leq \rho(A) + \varepsilon \ \text{ (Teil (1)}, \\
C^{-1}\|A\|_1 &\leq \|A\| \leq C\|A\|_1, \ \ C > 0 \ \text{(Äquivalenz zweier Normen)}, \\
U^H AU &= T, \ \ A = UTU^H, \ U^H U = I, \ T = (t_{ij}) \ \text{obere Dreiecksmatrix}, \\
&\qquad \rho(A) = \rho(T) = \max\{|t_{ii}|, \ i = 1, 2, ..., n\} \ \text{(Satz 2.15)}.
\end{aligned}$$

Nun führen wir eine Fallunterscheidung durch.

(3.1) Es sei zunächst $\rho(A) = 0$.
Wegen $0 = \rho(A) = \rho(T) = \max\{|t_{ii}|\}$ gilt $t_{ii} = 0, \ i = 1, 2, ..., n$, und T ist eine strikt obere Dreicksmatrix.
Die Matrixpotenz $T^m, \ m \geq 1$, hat somit in den ersten m Spalten ausschließlich Nullelemente, so dass spätestens die Matrix T^n eine Nullmatrix ist.
Damit berechnet man der Reihe nach

$$\begin{aligned}
A &= UTU^H, \\
A^2 &= UT\underbrace{U^H U}_{=I}TU^H = UT^2U^H, \\
A^3 &= UT^2\underbrace{U^H U}_{=I}TU^H = UT^3U^H, \\
&... \\
A^m &= UT^mU^H, \\
&... \\
A^n &= 0, \\
A^m &= 0 \ \text{ für } m \geq n.
\end{aligned}$$

Somit ist $\|A^m\| = 0$ für $m \geq n$ und $\lim\limits_{m\to\infty} \sqrt[m]{\|A^m\|} = 0 = \rho(A)$.

(3.2) Nun sei $\rho = \rho(A) > 0$ und $\lambda = \lambda(A)$.
Die Matrix $B = \frac{1}{\rho}A$ hat die EW $\frac{\lambda}{\rho}$.
Ihr betragsgrößter EW hat den Betrag Eins und $\rho(B) = 1$. Die neue These ist damit

$$1 = \lim_{m\to\infty} \sqrt[m]{\|B^m\|}.$$

Gemäß Teil (1) gibt es für B eine Matrixnorm $\|\cdot\|_\varepsilon$ mit $\varepsilon > 0$ und

$$\rho(B) \le \|B\|_\varepsilon \le \rho(B) + \varepsilon.$$

Daraus folgt für $\|B\|_\varepsilon$ und alle $m > 0$ die Ungleichungskette

$$1 = \rho(B) = \sqrt[m]{\rho(B)^m} = \sqrt[m]{\rho(B^m)} \le \sqrt[m]{\|B^m\|_\varepsilon} \le \sqrt[m]{\|B\|_\varepsilon^m} = \|B\|_\varepsilon \le \rho(B) + \varepsilon = 1 + \varepsilon$$

und

$$\limsup_{m\to\infty} \sqrt[m]{\|B^m\|_\varepsilon} \le 1 + \varepsilon.$$

Da diese Abschätzung für beliebiges $\varepsilon > 0$ richtig ist, folgt aus den beiden letzten Beziehungen

$$1 \le \liminf_{m\to\infty} \sqrt[m]{\|B^m\|_\varepsilon} \le \limsup_{m\to\infty} \sqrt[m]{\|B^m\|_\varepsilon} \le 1$$

(*limes inferior* $= \liminf_{m\to\infty} = \underline{\lim}$, *limes superior* $= \limsup_{m\to\infty} = \overline{\lim}$) und schließlich die Behauptung.

(3.3) Nun zeigen wir die Behauptung für eine beliebige Matrixnorm $\|\cdot\|$.
Aus der Normäquivalenz $C^{-1}\|S\|_\varepsilon \le \|S\| \le C\|S\|_\varepsilon$, $S = A^m$, und der Formel $\rho(A) = \lim_{m\to\infty} \sqrt[m]{\|A^m\|_\varepsilon}$ erhält man

$$\begin{aligned}
\rho(A) &= 1 \cdot \lim_{m\to\infty} \sqrt[m]{\|A^m\|_\varepsilon} = \lim_{m\to\infty} \sqrt[m]{C^{-1}} \lim_{m\to\infty} \sqrt[m]{\|A^m\|_\varepsilon} \\
&= \lim_{m\to\infty} \sqrt[m]{C^{-1}\,\|A^m\|_\varepsilon} \\
&\le \liminf_{m\to\infty} \sqrt[m]{\|A^m\|} \\
&\le \limsup_{m\to\infty} \sqrt[m]{\|A^m\|} \\
&\le \lim_{m\to\infty} \sqrt[m]{C\,\|A^m\|_\varepsilon} \\
&= \lim_{m\to\infty} \sqrt[m]{C} \lim_{m\to\infty} \sqrt[m]{\|A^m\|_\varepsilon} \\
&= \lim_{m\to\infty} \sqrt[m]{\|A^m\|_\varepsilon} \\
&= \rho(A)
\end{aligned}$$

und damit $\rho(A) = \lim_{m\to\infty} \sqrt[m]{\|A^m\|}$. □

Die Formel (2.57) ist die Basis für die Bestimmung von Konvergenzraten gemäß

$$R(A) = -\ln(\rho(A)) = \lim_{m\to\infty} R_m(A), \quad R_m(A) = -\ln\left(\sqrt[m]{\|A^m\|}\right) = -\tfrac{1}{m}\ln(\|A^m\|). \quad (2.58)$$

Der Satz ist mehr eine qualitative Aussage als ein konstruktives Verfahren. Auf Grund des Aufwands ihrer Berechnung sind leider weder der Spektralradius $\rho(A)$ noch die Norm $\|A\|_s$ praktikabel.

Nun kommen wir zur Konvergenz von Matrix-Folgen und -Reihen.
Wenn eine Matrixfolge $\{B_k\}$ gegen die Nullmatrix konvergiert, d. h. $\lim\limits_{k\to\infty} B_k = 0$ (also elementweise), dann ist das gleichbedeutend mit $\lim\limits_{k\to\infty} \|B_k\| = 0$ wegen der Definitheit der Norm. Man spricht in diesem Fall auch von einer Nullfolge.
Offenbar folgt aus $\lim\limits_{k\to\infty} B_k = B$ die Aussage $\lim\limits_{k\to\infty} \|B_k\| = \|B\|$. Das zeigt man gemäß

$$\begin{aligned}
0 &= \lim_{k\to\infty} (B_k - B), \\
0 &= \lim_{k\to\infty} \|B_k - B\|, \\
0 &= \lim_{k\to\infty} (-\|B - B_k\|), \\
&\quad -\|B - B_k\| \le \|B_k\| - \|B\| \le \|B_k - B\| \quad \text{(Dreiecksungleichung)}, \\
0 &= \lim_{k\to\infty} (\|B_k\| - \|B\|), \\
\|B\| &= \lim_{k\to\infty} \|B_k\|.
\end{aligned}$$

Die Umkehrung gilt nicht, weil natürlich eine ganz andere Matrix C, z. B. $-B$, dieselbe Norm wie B haben kann, jedoch sich von der Grenzmatrix eben unterscheiden kann.
Die Konvergenz der Matrixfolge notiert man auch als

$$\forall \varepsilon > 0 \; \exists k \in \mathbb{N} \; \forall m > k \quad \|B_m - B\| \; < \; \varepsilon. \tag{2.59}$$

Daraus ergibt sich die Cauchy-Bedingung

$$\forall \varepsilon > 0 \; \exists k \in \mathbb{N} \; \forall m > k \quad \|B_m - B_k\| \; < \; \varepsilon. \tag{2.60}$$

Bei Zahlenfolgen zeigt man die Äquivalenz der Formeln (2.59) und (2.60).
Die Konvergenz der Matrixreihe oder Neumannschen Reihe $I + A + A^2 + A^3 + ...$ heißt, dass die Summe

$$S = \sum_{k=0}^{\infty} A^k$$

existiert bzw. die Folge der Partialsummen $S_m = \sum\limits_{k=0}^{m} A^k$ konvergiert, d. h. $\lim\limits_{m\to\infty} S_m = S$.
Nach diesen Vorbetrachtungen formulieren wir den folgenden Satz, der für Zahlen mit der Konvergenz der geometrischen Reihe $\sum\limits_{k=0}^{\infty} q^k$ bei $|q| < 1$ korrespondiert.

Satz 2.53 *Folgende drei Aussagen über die Matrix A sind äquivalent:*
(1) Konvergenz der Matrixreihe (Neumannsche Reihe)

$$I + A + A^2 + A^3 + ..., \tag{2.61}$$

(2) Grenzwert

$$\lim_{m\to\infty} A^m \;=\; 0 \quad (\textit{Nullmatrix}), \tag{2.62}$$

(3) Spektralradius $\rho(A) < 1$.

Beweis. Wir verweisen zunächst auf die gerade gemachten Vorbetrachtungen.
Es genügt z. B., die Implikationskette (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1) $\Rightarrow$ zu zeigen.
(1) $\Rightarrow$ (2):
Mit der Konvergenz der Matrixreihe $S = \sum_{k=0}^{\infty} A^k$ ist die Summe S auch Grenzwert der Folge der Partialsummen $S_m = \sum_{k=0}^{m} A^k$, d. h. $\lim_{m\to\infty} S_m = S$.
Aber $A^m = S_m - S_{m-1}$ und durch den Grenzübergang folgt

$$\lim_{m\to\infty} A^m = \lim_{m\to\infty} (S_m - S_{m-1}) = \lim_{m\to\infty} S_m - \lim_{m\to\infty} S_{m-1} = S - S = 0.$$

(2) $\Rightarrow$ (3):
Sei $\lim_{m\to\infty} A^m = 0$, λ ein beliebiger EW von A und v mit $\|v\| = 1$ sein zugehöriger EV.
Die Größe λ^m ist somit EW von A^m. Wir zeigen, dass $|\lambda| < 1$ ist.
$\{A^m\}$ als Nullfolge heißt

$$\|A^m - 0\| \;<\; \varepsilon \;\text{ für alle } \varepsilon > 0 \text{ und } m > k,$$

und damit

$$\|\lambda^m v\| \;=\; \|(\lambda^m - 0)v\| \;=\; \|(A^m - 0)v\| \;\le\; \|A^m - 0\|\,\|v\| \;=\; \|A^m - 0\| \;<\; \varepsilon,$$
$$|\lambda|^m \;=\; |\lambda|^m\|v\| \;<\; \varepsilon,$$

was für beliebig kleines ε nur bei $|\lambda| < 1$ möglich ist. Also ist $\rho(A) = \max|\lambda(A)| < 1$.
(3) $\Rightarrow$ (1):
Nach Satz 2.52 folgt aus $\rho(A) < 1$ die Existenz einer Norm $\|A\| = s < 1$.
Daraus folgt wegen $\|A^m\| \le \|A\|^m = s^m \to 0$ für $m \to \infty$ der Grenzwert $\lim_{m\to\infty} A^m = 0$ (damit hat man zusätzlich den Nachweis der Implikation (3) $\Rightarrow$ (2)).
Weiter existieren die Inversen $(I \pm A)^{-1}$.
Wäre das nicht der Fall, so hätten die homogenen LGS $(I \pm A)x = 0$ eine nicht triviale

Lösung x, die der Beziehung $x = \mp Ax$ genügt. Mir den Normeigenschaften (Kompatibilitätsbedingung) $\|x\| = \|Ax\| \leq \|A\|\|x\|$ und nach Division durch $\|x\| \neq 0$ folgt weiter $1 \leq \|A\|$ im Widerspruch zu $\|A\| = s < 1$. Folglich haben die homogenen LGS $(I \pm A)x = 0$ nur die Nulllösung und die Matrizen $I \pm A$ sind regulär.
Die Partialsumme S_n erfüllt die Identität

$$\begin{aligned} (I-A)S_m &= I - A^{m+1} \\ \text{bzw. } (I-A)^{-1} - S_m &= (I-A)^{-1}A^{m+1}, \\ S_m &= (I-A)^{-1}(I-A^{m+1}). \end{aligned}$$

Nun kann man auf beiden Seiten wegen $\lim\limits_{m\to\infty} A^m = 0$ den Limes bilden und erhält

$$S = \lim_{m\to\infty} S_m = \sum_{k=0}^{\infty} A^k = (I-A)^{-1}.$$

Als Ergänzung zu den Thesen soll die Äquivalenzrelation (2) $\Leftrightarrow$ (3) wie in [15] gezeigt werden.
(2) $\Rightarrow$ (3):
Der Beweis ist indirekt. Angenommen, $\rho(A) \geq 1$. Dann gibt es zu A einen EW λ mit $|\lambda| \geq 1$ und einen zugehörigen EV $v \neq 0$ mit $Av = \lambda v$.
Wegen $A^m v = \lambda^m v$ und $\lim\limits_{m\to\infty} \lambda^m \neq 0$ kann $\{A^m\}$ daher keine Nullfolge sein.
(3) $\Rightarrow$ (2):
Hier brauchen wir wie in den Sätzen 2.32 und 2.52 die Jordan-Normalform der Matrix. Zunächst gilt mit einer regulären Matrix T die Beziehung

$$\lim_{k\to\infty} A^k = 0 \text{ gdw. } \lim_{k\to\infty} T^{-1}A^kT = 0.$$

Aber $T^{-1}A^kT = (T^{-1}AT)^k$, was leicht nachzuprüfen ist, so dass als Behauptung der Grenzwert $\lim\limits_{k\to\infty}(T^{-1}AT)^k = 0$ bleibt.
Die Matrix $A(n,n)$ lässt sich durch eine Ähnlichkeitstransformation auf die Jordan-Normalform $J = T^{-1}AT$ bringen. Wir zeigen nun, dass $\lim\limits_{k\to\infty} J^k = 0$ gilt, wenn alle EW λ_j von A dem Betrag nach kleiner als Eins sind. Dazu sei

$$J_j(t,t)(\lambda) = \begin{pmatrix} \lambda_j & 1 & & & \\ & \lambda_j & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_j \end{pmatrix}, \quad j = 1,2,...,p \leq n,$$

eine t-dimensionale Jordan-Zelle zum EW λ_j der Jordan-Normalform
$J = \text{diag}(J_1, J_2, ..., J_p)$. Da offenbar $J^k = \text{diag}(J_1^k, J_2^k, ..., J_p^k)$ gilt, genügt es, das Konvergenzverhalten der Jordan-Zellen J_j zu untersuchen.

Wir schreiben J_j in der zerlegten Form $J_j = \lambda_j I + S$ mit

$$S(t,t) \;=\; \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{pmatrix}$$

und bilden $J_j^k = (\lambda_j I + S)^k$. Nach Anwendung der binomischen Formel und unter Beachtung von $S^t = S^{t+1} = ... = S^n = S^{n+1} = ... = 0$ erhält man für großes k die Beziehung

$$J_j^k = \sum_{i=0}^{k} \binom{k}{i} \lambda_j^{k-i} S^i = \sum_{i=0}^{n-1} \binom{k}{i} \lambda_j^{k-i} S^i.$$

Für festes i hat man die Abschätzung

$$\left| \binom{k}{i} \lambda_j^{k-i} \right| = \left| \frac{k(k-1) \cdot ... \cdot (k-i+1)}{1 \cdot 2 \cdot ... \cdot i} \lambda_j^{k-i} \right| \leq |\lambda_j|^{k-i} k^i.$$

Aber die Zahlenfolge $\tau^{k-i} k^i,\ 0 \leq \tau < 1$, strebt für wachsendes k gegen Null, so dass damit und wegen $|\lambda_j| < 1$ die Konvergenz $\lim\limits_{k\to\infty} |\binom{k}{i} \lambda_j^{k-i}| = 0$ und weiter $\lim\limits_{k\to\infty} J_j^k = 0$ erfüllt ist. □

Wir treffen noch einige Aussagen über die Norm der Neumannschen Reihe (2.61).

Satz 2.54 Satz von Banach

Sei die Matrixnorm $\|A\| \leq s < 1$. Dann gelten die folgenden Aussagen.

(1) Die Reihe $I + A + A^2 + A^3 + ...$ ist konvergent und wir haben

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k \quad bzw. \quad I = (I - A) \sum_{k=0}^{\infty} A^k. \tag{2.63}$$

(2) Ist $\|\cdot\|$ eine Matrixnorm mit $\|I\| = 1$, so gilt die Normabschätzung

$$\frac{1}{1 + \|A\|} \leq \|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|} \leq \frac{1}{1 - s}. \tag{2.64}$$

(3) Genauso ist

$$(I + A)^{-1} = \sum_{k=0}^{\infty} (-A)^k \quad bzw. \quad I = (I + A) \sum_{k=0}^{\infty} (-A)^k \tag{2.65}$$

und

$$\frac{1}{1 + \|A\|} \leq \|(I + A)^{-1}\| \leq \frac{1}{1 - \|A\|}. \tag{2.66}$$

Beweis. Im Satz 2.53 liefert der Beweisteil (3) $\Rightarrow$ (1) die Existenz der Inversen $(I \pm A)^{-1}$ sowie die Summenformel

$$\sum_{k=0}^{\infty} A^k = (I - A)^{-1},$$

womit die Behauptung (1) gezeigt ist.
Analog ergibt sich aus Teil (3) die Gleichheit

$$\sum_{k=0}^{\infty} (-A)^k = (I + A)^{-1}.$$

Zum Nachweis der Normabschätzungen benutzen wir die beiden folgenden Identitäten mit jeweils der gleichen Vorzeichenwahl

$$I = (I \pm A)(I \pm A)^{-1}.$$

Mit Hilfe der Dreiecksungleichung und der Submultiplikativität der Matrixnorm erhält man

$$\begin{aligned}
1 \;=\; \|I\| &= \|(I \pm A)(I \pm A)^{-1}\| \\
&= \|(I \pm A)^{-1} \pm A(I \pm A)^{-1}\| \\
&\geq \|(I \pm A)^{-1}\| - \|A(I \pm A)^{-1}\| \\
&\geq \|(I \pm A)^{-1}\| - \|A\|\,\|(I \pm A)^{-1}\| \\
&= (1 - \|A\|)\,\|(I \pm A)^{-1}\|, \\
\|(I \pm A)^{-1}\| &\leq \frac{1}{1 - \|A\|} \;\leq\; \frac{1}{1 - s}
\end{aligned}$$

und

$$\begin{aligned}
1 \;=\; \|I\| &= \|(I \pm A)(I \pm A)^{-1}\| \\
&\leq \|(I \pm A)\|\,\|(I \pm A)^{-1}\| \\
&\leq (\|I\| + \|A\|)\|(I \pm A)^{-1}\| \\
&= (1 + \|A\|)\|(I \pm A)^{-1}\|, \\
\|(I \pm A)^{-1}\| &\geq \frac{1}{1 + \|A\|}.
\end{aligned}$$

□

Insbesondere dienen solche Normbetrachtungen für Fehlerabschätzungen der Näherungslösung von LGS.

2.4 Kondition

Den Begriff der Kondition brauchen wir im Zusammenhang mit der Empfindlichkeit (Sensitivität) von Problemen bzw. Algorithmen auf eingehende Fehler. Bei Matrizen oder LGS interessiert uns speziell das Verhalten von Verfahren zur Invertierung einer Matrix oder zur Lösung eines LGS bezüglich auftretender Störungen in der Matrix und/oder der rechten Seite des Systems.

2.4.1 Kondition und Konditionszahl einer Matrix

Definition 2.33 Konditionszahl

Für eine reguläre Matrix mit gegebener Norm sind die **(relative) Konditionszahl**

$$\text{cond}(A) \;=\; \|A\|\,\|A^{-1}\| \tag{2.67}$$

und die **absolute Konditionszahl**

$$\text{acond}(A) \;=\; \|A^{-1}\|. \tag{2.68}$$

Der Leser findet für die Konditionszahl auch die Bezeichnung $\kappa(A)$.

Es gilt wegen

$$1 \le \|I\| = \|AA^{-1}\| \le \|A\|\,\|A^{-1}\| \tag{2.69}$$

(bei induzierten Matrixnormen ist $1 = \|I\|$) die Beziehung $\text{cond}(A) \ge 1$.

Die Matrix ist schlecht konditioniert, falls $\text{cond}(A) \gg 1$. Ist A singulär, definiert man $\text{cond}(A) = \infty$. Allgemein gilt

$$\text{cond}(A) \;=\; \frac{\max\limits_{\|x\|=1} \|Ax\|}{\min\limits_{\|x\|=1} \|Ax\|}. \tag{2.70}$$

Wenn die Matrix fast singulär ist, also der Wert der Determinante nahe Null ist, dann ist mindestens ein EW fast Null. Ein weiterer EW liegt aber in der Größenordnung von $s_i = \sum\limits_j |a_{ij}|$. Sind letztere Summen groß, dann ist die Kondition von A im Allgemeinen schlecht. Das ist auch daran zu erkennen, dass für eine in ihren Elementen äquilibrierte (ausgewogene) Matrix A die Elemente der dazu inversen Matrix betragsmäßig groß werden. Die Kondition einer Matrix kann man somit als normunabhängig betrachten.

Will man die Abhängigkeit der Kondtion von einer speziellen Matrixnorm $\|A\|_s$ unterstreichen, so schreibt man diese als $\text{cond}_s(A)$.

Leider ist die Konditionszahl einer Matrix, Spezialfälle ausgenommen, nicht ohne größeren Aufwand berechenbar.

Für die Konditionszahl sowie für Genauigkeitsbetrachtungen brauchen wir die Kenntnis der Inversen von A oder wenigstens einer Näherungsinversen.
Eine gute Abschätzung für $\|A^{-1}\|$ und somit für $\text{cond}(A)$ ist oft rechnerisch aufwendig. Hat man jedoch die Faktorisierung von A, so kann man folgende effiziente Implementation der Inversenberechnung anwenden.
Wir beschränken uns auf die Notation der Inversenberechnung auf der Basis der Faktorisierung $A = LU$ ohne Pivotstrategie im Fall einer streng regulären Matrix A (siehe [8]). Man findet dort auch den allgemeinen Fall mittels der Faktorisierung einer regulären Matrix in der Form $PAQ = LU$ mit Totalpivotsuche sowie den Zeilen- bzw. Spaltenpermutationsmatrizen P und Q.

Algorithmus:
Seien $A = LU,\ L = (l_{ij}),\ l_{ii} = 1,\ U = (u_{ij})$ und $A^{-1} = (a'_{ij})$.
Rekursionsformeln zur Berechnung von a'_{ij}:

$$
\begin{aligned}
a'_{nn} &= \frac{1}{u_{nn}}, \\
k &= n-1, n-2, ..., 1 \\
&\left.\begin{aligned}
a'_{kj} &= -\frac{1}{u_{kk}} \sum_{i=k+1}^{n} u_{ki}\ a'_{ij} \\
a'_{jk} &= -\sum_{i=k+1}^{n} a'_{ji}\ l_{ik} \\
a'_{kk} &= \frac{1}{u_{kk}} - \sum_{i=k+1}^{n} a'_{ki}\ l_{ik}.
\end{aligned}\right\} \ j = k+1, k+2, ..., n,
\end{aligned}
$$

Beispiel 2.35 Sei $A = A(n,n) = (a_{ij})$ die Hilbert-Matrix. Ihre Elemente sind

$$a_{ij} = \frac{1}{i+j-1}.$$

Sie ist symmetrisch, positiv definit und ihre Inverse $A^{-1} = (a'_{ij})$ hat die ganzzahligen Elemente

$$a'_{ij} = \frac{(-1)^{i+j}}{i+j-1}\gamma_i\gamma_j, \quad \gamma_i = \frac{(n+i-1)!}{(i-1)!^2(n-i)!}, \quad i,j = 1,2,...,n.$$

Die Kondition von A verschlechtert sich mit wachsendem n.

Wir betrachten die Durchführung des verketteten Gauß-Algorithmus (VGA) mit Spaltenpivotisierung für die Ermittlung der Inversen und die Berechnung der Determinante $\det(A)$ bei wachsender Dimension n. Dabei tendieren die Pivotelemente (PE) gegen Null.
Zu Grunde liegen Implementierungen in Turbo Pascal (TP) mit den GPF *double* (64 Binärstellen, 15-16 Dezimalstellen der Mantisse) und *extended* (80 Binärstellen, 19-20 Dezimalstellen der Mantisse) sowie Rechnungen in Matlab (*double* Präzision) und Maple.

Generell wird sich bei numerischen Rechnungen die Anzahl der signifikanten Stellen im Ergebnis proportional zur wachsenden Dimension n verringern. So ist z. B. bei der Berechnung von $\det(A)$ mit TP *double* Präzision bei $n = 13$ höchsten noch die Größenordnung des Wertes verlässlich, bei $n = 16$ nicht einmal diese, was der Vergleich mit der exakten Auswertung im CAS Maple zeigt.

Die Konditionszahlen $\text{acond}_\infty(A)$ und $\text{cond}_\infty(A)$ können aus den folgenden Tabellen ermittelt werden.

Programm	n	7	10	13	16	19
TP	$\det(A)$	4.8E-25	2.2E-53	2.6E-92	1.8E-135	-3.9E-180
double	$\|A\|_\infty$	2.593	2.929	3.180	3.381	3.548
	$\|A^{-1}\|_\infty$	3.8E+8	1.2E+13	2.3E+17	5.7E+17	5.9E+17
TP	$\det(A)$	4.8E-25	2.2E-53	1.4E-92	-3.2E-141	1.1E-192
extended	$\|A\|_\infty$	2.593	2.929	3.180	3.381	3.548
	$\|A^{-1}\|_\infty$	3.8E+8	1.2E+13	4.2E+17	7.3E+20	7.5E+20
Matlab	$\det(A)$	4.8E-25	2.2E-53	1.4E-92	2.4E-135	-2.2E-180
double	$\|A\|_\infty$	2.593	2.929	3.180	3.381	3.548
	$\|A^{-1}\|_\infty$	3.8E+8	1.2E+13	1.1E+17 (a) 1.3E+17 (b)	6.8E+17	1.9E+18

Tab. 2.4 Berechnungen für die Hilbert-Matrix mit TP und Matlab, wobei
(a): `A\I`,
(b): `inv(A)` (ab $n \geq 13$ mit Warnungen)

Für die Dimensionen $n = 16$ und 19 sind die Berechnungen der Determinante und Kondition der Matrix A in TP bzw. Matlab mit der vorliegenden GPA nicht vertretbar.

Programm	n	7	10	13	16	19
Maple	$\det(A)$	4.8E-25	2.2E-53	1.4E-92	1.4E-142	2.0E-203
exakt	$\|A\|_\infty$	2.593	2.929	3.180	3.381	3.548
	$\|A^{-1}\|_\infty$	3.8E+8	1.2E+13	4.2E+17	1.5E+22	5.4E+26

Tab. 2.5 Berechnungen für die Hilbert-Matrix mit Maple

Rechnungen in Maple zu Norm und Kondition

Hilbert-Matrix A spd

```
> Digits:=16:
  i:='i': j:='j':
  n:=5:
  A:=matrix(n,n):
  A:=matrix(n,n,(i,j)->1/(i+j-1)):   # hilbert(n)
  `A`=evalm(A);
  `rank(A)`=rank(A);
  `det(A)`=det(A);
  `inv(A)`=inverse(A);

  # Normen
  `norm(A,2)`=norm(A,2);                      # Spektralnorm
  evalf(%);
  `norm(A,1)`=norm(A,1);                      # Spaltensummennorm
  evalf(%);
  `norm(A)=norm(A,infinity)`=norm(A);         # Zeilensummennorm
  evalf(%);
  `norm(A,frobenius)`=norm(A,frobenius);      # Frobenius-Norm
  evalf(%);

  # Konditionen dazu, auch Kommando cond(A,*)
  `cond(A,2)`=evalf(norm(A,2)*norm(inverse(A),2));
  `cond(A,1)`=evalf(norm(A,1)*norm(inverse(A),1));
  `cond(A)=cond(A,infinity)`=evalf(norm(A)*norm(inverse(A)));
  `cond(A,frobenius)`=evalf(norm(A,frobenius)*norm(inverse(A),frobenius));
```

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{bmatrix}$$

$$rank(A) = 5$$

$$det(A) = \frac{1}{266716800000}$$

$$inv(A) = \begin{bmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 1050 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{bmatrix}$$

$$norm(A,2) =$$

$$\frac{1}{5} RootOf(-1 + 3700542505\_Z - 1582832489513760\_Z^2 + 487666609069973760\_Z^3$$
$$-455148325561466880\_Z^4 + 7284515983589376\_Z^5, index = 5)^{1/2}$$

$$norm(A,2) = 1.567050691098231$$

$$norm(A,1) = \frac{137}{60}$$
$$norm(A,1) = 2.283333333333333$$
$$norm(A) = norm(A,infinity) = \frac{137}{60}$$
$$norm(A) = norm(A,infinity) = 2.283333333333333$$
$$norm(A,frobenius) = \frac{\sqrt{15871330}}{2520}$$
$$norm(A,frobenius) = 1.580906263272022$$
$$cond(A,2) = 476607.2502425608$$
$$cond(A,1) = 943656.$$
$$cond(A) = cond(A,infinity) = 943656.$$
$$cond(A,frobenius) = 480849.1169947188$$

Berechnung der EW (Informationen aus der Maple-Hilfe)

- **Eigenvals(A)** returns an array of the eigenvalues of A. The function Eigenvals itself is inert. To actually compute the eigenvalues and eigenvectors, the user must evaluate the inert function in the floating point domain, by **evalf(Eigenvals(A))**.
- The call **eigenvalues(A)** returns for a symbolic case a sequence of the eigenvalues of A computed by solving the characteristic polynomial $\det(\lambda I - A) = 0$ or for larger dimension (greater than four) the eigenvalues are expressed using Maple's **RootOf** notation for algebraic extensions. If A contains floating-point numbers, a numerical method is used where all arithmetic is done at the precision specified by **Digits**.

```
> charpoly(A,lambda);
  Eigenvals(A);
  evalf(Eigenvals(A));
  eigenvals(A);
  evalf(eigenvals(A));
```

$$\lambda^5 - \frac{563}{315}\lambda^4 + \frac{735781}{2116800}\lambda^3 - \frac{852401}{222264000}\lambda^2 + \frac{61501}{53343360000}\lambda - \frac{1}{266716800000}$$

$$\text{Eigenvals}\left(\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{bmatrix}\right)$$

$$[0.3287928771915000\,10^{-5},\ 0.0003058980401511738,\ 0.01140749162341977,$$
$$0.2085342186110133,\ 1.567050691098231]$$

$$\%1 := -1 + 61501\,\_Z - 40915248\,\_Z^2 + 741667248\,\_Z^3 - 762725376\,\_Z^4 + 85349376\,\_Z^5$$

$$\frac{1}{5}\,RootOf(\%1,\, index = 1),\ \frac{1}{5}\,RootOf(\%1,\, index = 2),\ \frac{1}{5}\,RootOf(\%1,\, index = 3),$$
$$\frac{1}{5}\,RootOf(\%1,\, index = 4),\ \frac{1}{5}\,RootOf(\%1,\, index = 5)$$

$$0.3287928772171862\,10^{-5},\ 0.0003058980401511918,\ 0.01140749162341981,$$
$$0.2085342186110134,\ 1.567050691098231$$

Prozedur für die Berechnung der Spektralnorm über EW von AA^T (bzw. A^TA)

```
> norm2:=proc(A::matrix)
        local n,i,B,EVB,seq_EVB;
   B:=evalm(A&*transpose(A));
   n:=linalg[rowdim](B);
   EVB:=evalf(Eigenvals(B));
   seq_EVB:=seq(EVB[i],i=1..n);
   sqrt(max(seq_EVB));
  end:

> norm2(A);
  evalf(norm(A,2));
  Digits:=10:
```

$$1.567050691098231$$
$$1.567050691098231$$

Kondition der Hilbert-Matrix abhängig von der Dimension n

```
> Digits:=16:
  i:='i': j:='j':
  printf(` n          cond(H(n,n))\n`):
  for n from 1 to 10 do
   A:=matrix(n,n,(i,j)->1/(i+j-1)):
   erg:=evalf(norm(A,2)*norm(inverse(A),2));
   printf(`%2d    %.15  \n`,n,erg);
  end do:
```

```
 n          cond(H(n,n))              Rundungseffekt
 1    1.000000000000000e+00
 2    1.928147006790397e+01
 3    5.240567775860608e+02
 4    1.551373873893259e+04   <-  1.551373873893258 8222...e+04
 5    4.766072502425608e+05
 6    1.495105864013122e+07
 7    4.753673549881789e+08   <-  4.753673549881789 9255...e+08
 8    1.525757574164694e+10
 9    4.931549269715421e+11
10    1.602628687021688e+13
```

Beispiel 2.36 Wir vergleichen gut und schlecht konditionierte LGS.
Ausgangspunkt ist die Version des LGS $Ax = b$, welche eine günstige Kondition der Koeffizientenmatrix besitzt.

$$\begin{pmatrix} \frac{1}{200} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}.$$

Die exakte Lösung ist $(\frac{5000}{9950}, \frac{4950}{9950})^T = (\frac{100}{199}, \frac{99}{199})^T$.
Die beiden anderen LGS mit derselben exakten Lösung entstehen einfach durch geeignete Zeilenmanipulationen des Systems. Wir betrachten ihre Auswirkungen auf die Eigenschaften der Koeffizientenmatrizen. Gleitpunktzahlen (GPZ) sind mit maximal 10 signifikanten Dezimalziffern notiert (Maple Format `Digits:=10`).

	1. LGS, $A = A^T$ $\begin{pmatrix} \frac{1}{200} & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$	2. LGS $\begin{pmatrix} 1 & 200 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 100 \\ 1 \end{pmatrix}$	3. LGS $\begin{pmatrix} \frac{1}{200} & 1 \\ \frac{19999}{100} & -200 \end{pmatrix}, \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$
A^{-1}	$\begin{pmatrix} -\frac{200}{199} & \frac{200}{199} \\ \frac{200}{199} & -\frac{1}{199} \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{199} & \frac{200}{199} \\ \frac{1}{199} & -\frac{1}{199} \end{pmatrix}$	$\begin{pmatrix} \frac{20000}{20099} & \frac{100}{20099} \\ \frac{19999}{20099} & -\frac{1}{40198} \end{pmatrix}$
$C = AA^T$	$\begin{pmatrix} \frac{40001}{40000} & \frac{201}{200} \\ \frac{201}{200} & 2 \end{pmatrix}$	$\begin{pmatrix} 40001 & 201 \\ 201 & 2 \end{pmatrix}$	$\begin{pmatrix} \frac{40001}{40000} & -\frac{3980001}{20000} \\ -\frac{3980001}{20000} & \frac{799960001}{10000} \end{pmatrix}$
$\lambda(A)$	-0.6144181931 1.619418194	15.14213562 -13.14213563	0.9999753 -200.9949752
$\lambda(C)$	0.3775097159 2.622515284	0.98998 40002.01002	0.5049843749 79996.49513
$p_2(\lambda(A))$	$\lambda^2 - \frac{201}{200}\lambda - \frac{199}{200}$	$\lambda^2 - 2\lambda - 199$	$\lambda^2 - \frac{39999}{200}\lambda - \frac{20099}{100}$
$p_2(\lambda(C))$	$\lambda^2 - \frac{120001}{40000}\lambda + \frac{39601}{40000}$	$\lambda^2 - 40003\lambda + 39601$	$\lambda^2 - \frac{639976001}{8000}\lambda + \frac{403969801}{10000}$
$\|A\|_2$	1.619418193	200.0050250	282.8365166
$\|A^{-1}\|_2$	1.627555973	1.005050377	1.407216860
$\text{cond}_2(A)$	2.635693753	201.0151258	398.0123148
$\text{cond}_\infty(A)$	4.020100503	203.0201005	400.0099010
$\text{cond}_1(A)$	4.020100503	203.0201005	400.0099010

Anweisungen in Maple zu EW, Norm und Kondition

```
> Digits:=10:

> A:=matrix(2,2,[[1/200,1],
                 [1, 1]]);
  B:=inverse(A);
  C:=evalm(A&*transpose(A));

> b:=vector(2,[1/2,1]);
  linsolve(A,b);

> # EW
  vecsan:='vecsan':
  lambdan:=evalf(Eigenvals(A,vecsan));
  eigenvals(A);
  vecsan:='vecsan':
  lambdan:=evalf(Eigenvals(C,vecsan));
  eigenvals(C);

> # Charakterisches Polynom
  charpoly(A,lambda);
  charpoly(C,lambda);

> # Spektralnorm
  norm(A,2); norm2(A);
  norm(B,2); norm2(B);

> # Spektralkondition
  nc2:=evalf(norm2(A)*norm2(B));
  nc2_err:=evalf(norm(A,2)*norm(B,2));
> # ZS-Kondition
  ncinf:=evalf(norm(A,infinity)*norm(B,infinity));
> # SS-Kondition
  nc1:=evalf(norm(A,1)*norm(B,1));
```

Zunächst sei bemerkt, dass für eine gute Kondition die Symmetrie der Koeffizientenmatrix eine vorteilhafte Eigenschaft darstellt. Wichtiger jedoch ist die Ausgewogenheit der Matrix, d. h. dass die Zeilensummen $\sum_j |a_{ij}|$ und Spaltensummen $\sum_i |a_{ij}|$ ungefähr von gleicher Größenordnung sind. Die Matrixskalierung dient diesem Ziel. Dazu kommt noch, dass möglichst viele Matrixelemente von gleicher Größenordnung sein sollen.
In diesem Beispiel ist das bei der Matrix des ersten LGS im Vergleich mit den anderen Matrizen der Fall.

Beispiel 2.37 Betrachten wir noch das LGS

$$\begin{pmatrix} 2 & 6 \\ 2 & 6.00001 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 8 \\ 8.00001 \end{pmatrix}.$$

Die exakte Lösung ist $(1,1)^T$.
Die Inverse der Koeffizientenmatrix A ist

$$A^{-1} = \begin{pmatrix} 300000.5 & -300000 \\ -100000 & 100000 \end{pmatrix}.$$

Weiterhin haben wir $\sigma(A) = \{0.2500 \cdot 10^{-5},\ 8.000007500\}$.
Für die Matrix

$$C = AA^T = \begin{pmatrix} 40 & \frac{2000003}{50000} \\ \frac{2000003}{50000} & \frac{400001200001}{10000000000} \end{pmatrix} = \begin{pmatrix} 40 & 40.000006 \\ 40.000006 & 40.0001200001 \end{pmatrix}$$

liefert Maple mit der Genauigkeit `Digits=10` das Spektrum $\sigma(C) = \{0.,\ 80.00012001\}$.
D. h. jedoch, dass das Format nicht ausreicht, denn der EW Null ist natürlich eine numerische Näherung. Mit `Digits=20` erhält das Spektrum

$$\sigma(C) = \{0.4999993 \cdot 10^{-11},\ 80.000120000095000009\}.$$

Wir haben also die Situation einer fast singulären Matrix A mit einem EW fast Null. Durch das Matrixprodukt $C = AA^T$ wird die Kondition eher schlechter. Der EW nahe Null wird noch kleiner. Ein weiterer EW liegt aber in der Größenordnung von $\sum_j |c_{ij}|$.
Die Konditionen der Matrix A sind

$$\mathrm{cond}_2(A) = 0.4000006001 \cdot 10^7, \quad \mathrm{cond}_\infty(A) = \mathrm{cond}_1(A) = 0.4800010000 \cdot 10^7.$$

Bei ausgewogenen Matrixelementen von A ist ihre schlechte Kondition daran zu erkennen, dass die Elemente der dazu inversen Matrix betragsmäßig groß werden.
Die äquilibrierte Form von A muss man schon voraussetzen, denn für die mit einem großen Faktor durchmultiplizierte Matrix

$$\widetilde{A} = \begin{pmatrix} 2 \cdot 10^5 & 6 \cdot 10^5 \\ 2 \cdot 10^5 & 6.00001 \cdot 10^5 \end{pmatrix}$$

würde man

$$\widetilde{A}^{-1} = (a'_{ij}) = \begin{pmatrix} 3.000005 & -3 \\ -3 & 1 \end{pmatrix}$$

mit $a'_{ij} = \mathcal{O}(1)$ erhalten.

Dem Leser überlassen wir die analoge Untersuchung des LGS

$$\begin{pmatrix} 2 & 6 \\ 2 & 5.99999 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 8 \\ 8.00002 \end{pmatrix}$$

mit der exakten Lösung $(10,-2)^T$.

2.4.2 Eigenschaften der Konditionszahl

Es sollen nur einige spezielle Eigenschaften der Konditionszahl aufgelistet werden.

1. Ist $A = \mathrm{diag}(d_1, d_2, ..., d_n),\ d_i \neq 0$, so lautet die Konditionszahl mit $\|A\|_\infty$

$$\mathrm{cond}_\infty(A) \ = \ \frac{\max\limits_{i=1(1)n} |d_i|}{\min\limits_{i=1(1)n} |d_i|}.$$

2. Mit $0 < \varepsilon \ll 1$ ist die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1+\varepsilon & 1-\varepsilon \end{pmatrix}$$

fast singulär, und es folgt in der Norm $\|A\|_\infty$ die Konditionszahl

$$\mathrm{cond}_\infty(A) \ = \ 1 + \frac{2}{\varepsilon}.$$

3. Möglichkeiten der Berechnung der spektralen Konditionszahl sind

$$\mathrm{cond}_2(A) \ = \ \begin{cases} \|A\|_2 \ \|A^{-1}\|_2 \ = \ \dfrac{\max\limits_{i=1(1)n} \sqrt{\lambda_i(A^T A)}}{\min\limits_{i=1(1)n} \sqrt{\lambda_i(A^T A)}}, & \text{falls } A \text{ regulär}, \\ \dfrac{\max\limits_{i=1(1)n} |\lambda_i(A)|}{\min\limits_{i=1(1)n} |\lambda_i(A)|}, & \text{falls } A = A^T \text{ regulär}, \\ \dfrac{\max\limits_{i=1(1)n} \lambda_i(A)}{\min\limits_{i=1(1)n} \lambda_i(A)}, & \text{falls } A = A^T > 0 \ (A \text{ spd}). \end{cases}$$

Die Größen $\sigma_i = \sqrt{\lambda_i(A^T A)} \geq 0$ werden Singulärwerte der Matrix A genannt.

4. Weitere Merkmale sind im folgenden Satz angegeben.

Satz 2.55 *Für die jeweils genannten Matrizen, Konstanten und Konditionszahlen (mit dem mehrfachen Index sind ihre verschiedenen normabhängigen Definitionen bezeichnet) gelten die folgenden Aussagen.*

(a) $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\ \mathrm{cond}(B)$ *für alle Matrixnormen.*

(b) $\mathrm{cond}(cA) = \mathrm{cond}(A)$ *für alle* $c \in \mathbb{R},\ c \neq 0$.

(c) $\mathrm{cond}_2(Q) = 1$ *für Q orthogonal.*

(d) $\mathrm{cond}_2(A) \leq \mathrm{cond}_F(A) \leq \mathrm{cond}_G(A) \leq n^2\, \mathrm{cond}_\infty(A)$.

(e) $n^{-1}\, \mathrm{cond}_{1,F,\infty}(A) \leq \mathrm{cond}_2(A) \leq n\, \mathrm{cond}_{1,\infty}(A)$.

(f) $\mathrm{cond}_2(QA) = \mathrm{cond}_2(A)$ *für Q orthogonal.*

2.4.3 Schätzungen der Konditionszahl

Als grobe Näherung der Konditionszahl gilt bei einer Dreiecksfaktorisierung $A = LU$ ($l_{ii} = 1$) die aus den Diagonalelementen u_{ii} gebildete Zahl

$$\text{cond}_N(A) \;=\; \max_{1\le i,j\le n} \frac{|u_{ii}|}{|u_{jj}|}. \tag{2.71}$$

Man kann diese Beziehung plausibel machen mittels der Betrachtung der QR-Faktorisierung von A mit der Orthogonalmatrix Q und oberen Dreiecksmatrix R, sowie mit der dann geltenden Beziehung

$$\text{cond}_2(A) \;=\; \text{cond}_2(QR) \;=\; \text{cond}_2(R) \;\approx\; \frac{\max|\lambda(R)|}{\min|\lambda(R)|} \;=\; \max_{i,j}\frac{|r_{ii}|}{|r_{jj}|}. \tag{2.72}$$

Für Diagonalmatrizen stimmt $\text{cond}_N(A)$ mit den Konditionszahlen $\text{cond}_\infty(A)$ und $\text{cond}_1(A)$ überein.

Im Rahmen von Gleichungssystemlösern sollen hier noch drei Konditionsschätzungen angegeben werden, die ebenfalls auf der LU-Faktorisierung von A basieren [7].

(1) Hadamardsche **Konditionszahl**.

$$\text{hcond}(A) \;=\; \prod_{i=1}^{n}\frac{|r_{ii}|}{\|a_i\|_2} \;=\; \frac{|\det(A)|}{\prod_{i=1}^{n}\|a_i\|_2} \;\le\; 1, \tag{2.73}$$

mit a_i als Zeilenvektoren von A.
Ist der Wert sehr viel kleiner als 1, so ist die Matrix schlecht konditioniert.

(2) Konditionsschätzung nach Forsythe/Moler.
Dazu braucht man die durch den Gauß-Algorithmus ermittelte Lösung x_0 des LGS $Ax_0 = b$ mit beliebiger rechter Seite (z. B. Einsvektor), das mit doppelter Genauigkeit berechnete Residuum $r_0 = b - Ax_0$ sowie die Lösung des Gleichungssystems für den Fehler $Az = r_0$ unter Verwendung der schon durchgeführten Dreiecksfaktorisierung. Dann erhält man als Schätzung

$$\text{fcond}(A) \;=\; 2^t\frac{\|z\|_2}{\|x_0\|_2} \;\in\; (0,\infty) \tag{2.74}$$

mit der Maschinengenauigkeit 2^{-t}. Beim GPF *double* Präzision beträgt diese Größe $2^{-52} \approx$ 2.2E-16.

Wird dieser Wert sehr groß im Vergleich zur Eins, so kann man die Matrix als schlecht konditioniert betrachten. Werte um bzw. kleiner als Eins verweisen auf eine gute Kondition.

(3) Konditionsschätzung nach CLINE.

Hier wird die Kondition $\|A\|_\infty \|A^{-1}\|_\infty$ abgeschätzt. Dazu braucht man die Faktorisierung $PA = LU$ mit der Permutationsmatrix P.
Für U^T müssen dann $x = (\pm 1, \pm 1, ..., \pm 1)^T$ und $y = U^{-T}x$ so bestimmt werden, dass $\|y\|_\infty$ oder $\|y\|_1$ möglichst groß wird.
Weiter ist durch Rückwärtselimination das Gleichungssystem $L^T z = y$ zu lösen. Somit erhält man die Näherung $K = \|z\|_\infty / \|x\|_\infty$ für $\|A^{-1}\|_\infty$. Noch besser ist jedoch der Schätzwert $K = \|z\|_2 / \|x\|_2$.
Der Schätzwert für die Kondition ergibt sich als $\|A\|K$.

Der größte Aufwand liegt hier in der Bestimmung der Vektoren x, y, der als Programmausschnitt in TP nachfolgend angegeben ist.

```
x[1]:=1;
y[1]:=1/U[1,1];            { Komponenten x[1], y[1] fest }
                           { rechte Seite zunaechst = 0 }
for i:=2 to n do
 y[i]:=-R[1,i]*y[1]/U[i,i];

for k:=2 to n do
 begin
  v:=1/U[k,k];         { Beruecksichtigung von +-1 }
  x[k]:=y[k]-v;        { x[k..n] gleichzeitig als Hilfsvektor }
  y[k]:=y[k]+v;
  SMI:=abs(x[k]);
  SPL:=abs(y[k]);

  for i:=k+1 to n do
   begin
    v:=U[k,i]/U[i,i];
    x[i]:=y[i]-v*x[k];   y[i]:=y[i]-v*y[k];
    SMI:=SMI+abs(x[i]);  SPL:=SPL+abs(y[i]);
   end;

  if SMI>SPL then begin
                   for i:=k to n do y[i]:=x[i];
                   x[k]:=-1;
                  end
             else x[k]:=1;
 end;
```

2.4.4 Konditionszahl und Lösung von LGS

Sei $Ax = b$ das **exakte LGS** mit der exakten Lösung x.

Das **gestörte LGS** $\tilde{A}\tilde{x} = \tilde{b}$ enthalte die folgenden Größen:

$$\begin{aligned} \tilde{A} &= A + \Delta A \quad \text{gestörte Matrix,} \\ \tilde{b} &= b + \Delta b \quad \text{gestörte rechte Seite,} \\ \tilde{x} &= x + \Delta x \quad \text{Näherungslösung, gestörte Lösung.} \end{aligned} \tag{2.75}$$

Die Interpretation der fehlerbehafteten Größen sei:

A "Eingangsmatrix" mit "Eingangsstörungen" ΔA,

b "Eingangsvektor" mit "Eingangsstörungen" Δb,

x "Ausgangsvektor" mit "Ausgangsfehler" Δx.

Wir untersuchen nun die Frage, wie sich die relative Fehler $\|\Delta A\|/\|A\|$ und $\|\Delta b\|/\|b\|$ der Eingangsgrößen auf den relativen Fehler $\|\Delta x\|/\|x\|$ der Lösung auswirkt.

In der Antwort darauf werden auch hier die Kondition und Konditionszahlen $\mathrm{cond}(A)$ bzw. $\mathrm{acond}(A)$ der Matrix A eine wichtige Rolle spielen.

Die Konditionszahl dient zur Fehlerschätzung der Lösung von LGS.

Satz 2.56 *Ist A regulär und gilt die vereinfachende Annahme $\Delta A = 0$, so erhält man mit beliebiger gegebener Norm $\|x\|$ und induzierter Matrixnorm $\|A\|$ die a-priori-Fehlerschätzung*

$$\frac{\|\Delta x\|}{\|x\|} \;\le\; \mathrm{cond}(A)\,\frac{\|\Delta b\|}{\|b\|}. \tag{2.76}$$

Beweis. (Skizze)

Es gelten die Beziehungen

$$Ax = b \qquad \Rightarrow \|b\| = \|Ax\| \le \|A\|\,\|x\| \quad \text{und} \quad \|x\| \ge \frac{\|b\|}{\|A\|},$$

$$A(x + \Delta x) = b + \Delta b \quad \Rightarrow \Delta x = A^{-1}\Delta b \qquad \text{und} \quad \|\Delta x\| \le \|A^{-1}\|\,\|\Delta b\|.$$

Aus beiden Ungleichungen folgt die Behauptung

$$\frac{\|\Delta x\|}{\|x\|} \;\le\; \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|} \;=\; \mathrm{cond}(A)\frac{\|\Delta b\|}{\|b\|}.$$

□

Satz 2.57 *Ist A regulär, $\|\Delta A\|$ hinreichend klein und gilt die vereinfachende Annahme $\Delta b = 0$, so erhält man mit beliebiger gegebener Norm $\|x\|$ und induzierter Matrixnorm $\|A\|$ die a-priori-Fehlerschätzung*

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \frac{\|\Delta A\|}{\|A\|}. \tag{2.77}$$

Beweis. Wegen A regulär und hinreichend kleiner Störungen der Matrixkoeffizienten setzen wir auch die Regularität von $A + \Delta A$ voraus. Aus

$$Ax = b \quad \text{und} \quad (A + \Delta A)(x + \Delta x) = b$$

folgen

$$\begin{aligned} (A + \Delta A)x + (A + \Delta A)\Delta x &= b \\ (A + \Delta A)\Delta x &= -\Delta A\, x \\ \Delta x &= -(A + \Delta A)^{-1}\Delta A\, x \\ &= -[A(I + A^{-1}\Delta A)]^{-1}\Delta A\, x \\ &= -(I + A^{-1}\Delta A)^{-1}\, A^{-1}\Delta A\, x. \end{aligned} \tag{2.78}$$

Die Inverse von $I + A^{-1}\Delta A$ existiert also und wir erhalten dies auch durch die einfache zulässige Annahme $\|A^{-1}\Delta A\| \leq \|A^{-1}\|\,\|\Delta A\| < 1$.
Weitere Umformungen und Abschätzungen ergeben

$$\begin{aligned} \Delta x &= -GFx, \quad F = A^{-1}\Delta A,\ G = (I + F)^{-1}, \\ \|\Delta x\| &\leq \|G\|\,\|F\|\,\|x\|, \quad \|G\| \leq \frac{1}{1 - \|F\|} \text{ wegen Lemma 2.60}, \\ \frac{\|\Delta x\|}{\|x\|} &\leq \frac{\|F\|}{1 - \|F\|}, \quad \|F\| \leq \|A^{-1}\|\,\|\Delta A\| \\ &\leq \frac{\|A^{-1}\|\,\|\Delta A\|}{1 - \|A^{-1}\|\,\|\Delta A\|} \\ &\leq \frac{\|A\|\,\|A^{-1}\|\,\frac{\|\Delta A\|}{\|A\|}}{1 - \|A\|\,\|A^{-1}\|\,\frac{\|\Delta A\|}{\|A\|}} \\ &\leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\,\frac{\|\Delta A\|}{\|A\|}} \frac{\|\Delta A\|}{\|A\|}. \end{aligned}$$

□

Satz 2.58 *Hat man es bei der Lösung von $Ax = b$ neben Fehlern in b und x zusätzlich mit (hinreichend kleinen) Störungen der Matrixkoeffizienten $\Delta A \neq 0$ zu tun, so gelten die Abschätzungen*

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right) \tag{2.79}$$

bzw.

$$\|\Delta x\| \leq \frac{\operatorname{acond}(A)}{1 - \operatorname{acond}(A)\|\Delta A\|} \left(\|\Delta A\| \, \|x\| + \|\Delta b\| \right). \tag{2.80}$$

Beweis. [19], [42]
Die Beweistechnik lehnt sich mit den gleichen Voraussetzungen an die von Satz 2.57 an.
Wir zeigen die Abschätzung (2.79), wobei als Zwischenschritt die Ungleichung (2.80) zu erkennen ist.
Aus

$$Ax = b \;\text{ und }\; (A + \Delta A)(x + \Delta x) = b + \Delta b$$

folgen

$$\begin{aligned}
\Delta x &= (I + A^{-1}\Delta A)^{-1} \left(-A^{-1}\Delta A \, x + A^{-1}\Delta b \right) \\
&= (I + F)^{-1}(-Fx + A^{-1}\Delta b), \quad F = A^{-1}\Delta A, \\
\|\Delta x\| &\leq \|(I + F)^{-1}\| \, \| - Fx + A^{-1}\Delta b\|, \quad \|(I + F)^{-1}\| \leq 1/(1 - \|F\|) \\
&\leq \frac{\|F\| \, \|x\| + \|A^{-1}\| \, \|\Delta b\|}{1 - \|F\|}, \quad \|F\| \leq \|A^{-1}\| \, \|\Delta A\| < 1 \\
&\leq \frac{\|A^{-1}\| \, \|\Delta A\| \, \|x\| + \|A^{-1}\| \, \|\Delta b\|}{1 - \|A^{-1}\| \, \|\Delta A\|}, \quad \text{Ergebnis (2.80)} \\
&\leq \frac{\|A^{-1}\| \, \|\Delta A\| \, \|x\| + \|A\| \, \|A^{-1}\| \, \frac{\|\Delta b\|}{\|A\|}}{1 - \|A^{-1}\| \, \|\Delta A\|}, \quad \|A\| \geq \|b\|/\|x\| \\
&\leq \frac{\|A\| \, \|A^{-1}\| \, \frac{\|\Delta A\|}{\|A\|} \, \|x\| + \|A\| \, \|A^{-1}\| \, \frac{\|\Delta b\|}{\|b\|} \, \|x\|}{1 - \|A\| \, \|A^{-1}\| \, \frac{\|\Delta A\|}{\|A\|}}, \\
\frac{\|\Delta x\|}{\|x\|} &\leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right).
\end{aligned}$$

□

Die Konsequenzen der Abschätzungen (2.79) und (2.80) sind zunächst, dass der Nenner der ersten Brüche der rechten Seite wohl definiert sein muss. Das heißt, das eine schlechte absolute oder relative Konditionszahl prinzipiell nur kleine Störungen der Matrix zulässt. Dann haben wir die praktische Bedeutung in numerischen Berechnungen bei einer d-stelligen dezimalen GPA.
Ist die Konditionszahl $\text{cond}(A) \approx 10^{\alpha}$ und $\nu = 5 \cdot 10^{-d}$, d Stellenanzahl der Rechnung, so ergibt sich mit

$$\nu \ \text{cond}(A) = 5 \cdot 10^{\alpha - d} \ll 1,$$

$$\|\Delta A\|/\|A\| \leq \nu \ \text{ und } \ \|\Delta b\|/\|b\| \leq \nu$$

aus der Beziehung (2.79) die qualitative Abschätzung

$$\frac{\|\Delta x\|}{\|x\|} \leq \text{cond}(A)\ 2\nu = 10^{-d+\alpha+1} = \varepsilon. \tag{2.81}$$

Bemerkung 2.5

1. Die Schätzung (2.81) sagt, dass bei Lösung eines LGS mit den obigen Annahmen auf Grund von Eingangsfehlern in der berechneten Lösung $\tilde{x}$ nur $d-\alpha-1$ Dezimalstellen, bezogen auf die betragsgrößte Komponente, sicher sind. Eine pessimistische Aussage, die aber eintreten kann.

2. Des Weiteren spielt neben der Störmatrix auch die Größe $s = d-\alpha-1$ bei einer Nachiteration des LGS eine Rolle. Nur wenn $s > 0$ ist, wird mit jedem Nachiterationsschritt wenigstens eine Verbesserung der Näherungslösung um s Dezimalstellen eintreten und die Folge der Näherungslösungen konvergiert tatsächlich. Die Berechnung des Residuums mit höherer Genauigkeit geht also konform mit der Vergrößerung der Mantissenlänge d (damit ist $s > 0$).

3. Die Abschätzung

   $$\|\Delta x\|_\infty/\|x\|_\infty \leq \varepsilon$$

   in der Maximumnorm garantiert nur, dass die betragsgroßen Komponenten von x einen durch ε beschränkten relativen Fehler haben. Der relative Fehler der betragskleinen Komponenten kann beliebig größer als ε sein. Feinere komponentenweise Abschätzungen der Form

   $$|\Delta x_i|/|x_i| \leq \varepsilon_i$$

   sind wesentlich komplizierter.

Abschließend noch die Betrachtung des GA mit der Akkumulation von Rundungsfehlern in den n Schritten. Die Rundungsfehleranalyse (siehe [19]) zeigt wiederum den Einfluss der Matrixkondition. Der erzeugte Fehler $\Delta x = \tilde{x} - x$ kann abgeschätzt

werden gemäß

$$\|\Delta x\| \;\leq\; \nu\; K(n)\; \mathrm{cond}(A)\; \|x\|, \tag{2.82}$$

wobei $\nu = 10^{-d}$ (bzw. 2^{-t}) das Fehlerniveau (Genauigkeit der GPA) und $K(n)$ die vom Verfahren und der Dimension n abhängige Kumulationskonstante bedeuten. Für den GA gilt grob

$$K(n) = \begin{cases} \mathcal{O}(n) & \text{ohne Pivotisierung bei} \\ & A \text{ diagonaldominant oder } A = A^T > 0, \\ \mathcal{O}(2^n) & \text{bei partieller Pivotisierung,} \\ \mathcal{O}(n^{3/2}) & \text{bei vollständiger Pivotisierung.} \end{cases} \tag{2.83}$$

Auf Grund der Normäquivalenz gelten die Abschätzungen (2.76), (2.77), (2.79) und (2.80) auch für kompatible Normen.

2.4.5 Fehlerschätzungen mit Rückwärtsanalyse für LGS

Es gibt eine Fülle von Varianten der Genauigkeitsbewertung von Ergebnissen der numerischen Lösung des LGS $Ax = b$. Sie beinhalten Abschätzungen für den Residuenvektor (Bildvektordifferenz) $r = Ax - b$ bzw. Forderungen an den absoluten oder relativen Fehler des Lösungsvektors. Letzteres passiert im Rahmen der sogenannten "Rückwärtsanalyse" (backward analysis), wo $x + \Delta x$ als exakte Lösung des leicht gestörten Systems

$$(A + \Delta A)x' = b + \Delta b \tag{2.84}$$

betrachtet wird. Dazu benutzen wir induzierte Matrixnormen.

Satz 2.59 *Für den Urbildfehler $\Delta x = \tilde{x} - x$ der Näherung $\tilde{x}$ der Lösung des LGS $Ax = b$ gilt die a-posteriori-Fehlerschätzung*

$$\|\Delta x\| \;\leq\; \|A^{-1}\|\,\|r\| \;=\; \mathrm{acond}(A)\,\|r\| \;\leq\; \frac{\|C\|}{1 - \|CA - I\|}\,\|r\|, \tag{2.85}$$

wobei C die genäherte Inverse von A mit $\|CA - I\| < 1$ ist und $S = CA - I$ die zu C gehörige Störmatrix (Restmatrix) darstellt.

Für den Nachweis dieses Satzes verwenden wir das folgende Störungslemma.

Lemma 2.60 Störungslemma

Ist $\|S\| < 1$ und I die Einheitsmatrix, so existieren die Matrizen $(I \pm S)^{-1}$ und es gilt

$$\frac{1}{1 + \|S\|} \;\leq\; \|(I \pm S)^{-1}\| \;\leq\; \frac{1}{1 - \|S\|}. \tag{2.86}$$

Beweis. Lemma 2.60 (Skizze)

(1) Regularität von $I + S$.

$x \neq 0, \quad \|(I+S)x\| = \|Ix+Sx\| \geq \|x\| - \|Sx\| \geq \|x\| - \|S\|\|x\| = (1-\|S\|)\|x\| > 0.$

(2) Sei $T = (I+S)^{-1}, \quad \|T\| \neq 0.$

$$1 = \|I\| = \|T^{-1}T\| = \|(I+S)T\| \geq \|T\| - \|S\|\|T\| = (1-\|S\|)\|T\| > 0$$
$$\Rightarrow \|T\| \leq \frac{1}{1-\|S\|}.$$
$$1 = \|I\| = \|T^{-1}T\| = \|(I+S)T\| \leq \|I+S\|\|T\| \leq (1+\|S\|)\|T\|$$
$$\Rightarrow \|T\| \geq \frac{1}{1+\|S\|}.$$

Analog ist der Nachweis für $-S$, indem S durch $-S$ ersetzt wird. □

Beweis. Satz 2.59 (Skizze)

Die Größe des **Residuenvektors** $r = A\tilde{x} - b$ der Näherung $\tilde{x}$ führt über die Gleichung $\Delta x = \tilde{x} - x = A^{-1}r$ und Normabschätzungen auf den ersten Teil der Ungleichungskette (2.85)

$$\|\Delta x\| \;\leq\; \|A^{-1}\|\,\|r\| \;=\; \mathrm{acond}(A)\|r\|.$$

Für den zweiten Teil dieser Ungleichung brauchen wir das Lemma 2.60. Damit gilt für die Inverse und den Fehler der Inversen

$$\|A^{-1}\| = \|(CA)^{-1}C\| = \|(I+S)^{-1}C\| \leq \|(I+S)^{-1}\|\,\|C\| \leq \frac{\|C\|}{1-\|S\|}.$$ □

Folgerung 2.61 *(1) Die genäherte Inverse C genügt der Abschätzung*

$$\|C-A^{-1}\| = \|(CA-I)A^{-1}\| \leq \|S\|\,\|A^{-1}\| \leq \frac{\|S\|\,\|C\|}{1-\|S\|},$$

also

$$\frac{\|C-A^{-1}\|}{\|C\|} \;\leq\; \frac{\|S\|}{1-\|S\|}. \tag{2.87}$$

Ist die Störmatrixnorm $\|S\| > 1$, so ist die Berechnung von C offenbar mit so großen Fehlern behaftet, dass diese mit numerisch stabileren Algorithmen und/oder erhöhter arithmetischer Genauigkeit notwendig ist.

(2) Mit der Beziehung (2.85) und $\|b\| \leq \|A\|\|x\|$ erhält man unmittelbar

$$\frac{\|\Delta x\|}{\|x\|} \;\leq\; \|A\|\,\|A^{-1}\|\frac{\|r\|}{\|b\|} \;=\; \mathrm{cond}(A)\,\frac{\|r\|}{\|b\|}. \tag{2.88}$$

(3) Beide Ungleichungen (2.85) und (2.88) sind Abschätzungen für den absoluten bzw. relativen Fehler. Sie bedeuten, dass neben einem kleinen Residuenvektor r die Elemente der inversen Matrix bzw. die Kondition der Matrix Ausschlag gebend für den Fehler der Näherung $\tilde{x}$ sind. Ein kleines Residuum korrespondiert nicht automatisch mit einem kleinen Lösungsfehler.

(4) Betrachtet man also den Einfluss von Störungen der rechten Seite gemäß der Beziehung $A(x + \Delta x) = b + \Delta b$, so sind $A\Delta x = \Delta b$ und

$$\|\Delta x\| \;\leq\; \|A^{-1}\|\,\|\Delta b\| \;\leq\; \frac{\|C\|}{1 - \|S\|}\|\Delta b\| \tag{2.89}$$

mit der Näherungsinversen C und der Störmatrix $S = CA - I$.
Im Prinzip ist es eine andere Interpretation der Beziehung (2.85), wobei das Residuum r durch die Störung Δb ersetzt worden ist.

Beispiel 2.38 Gesucht sei die Lösung des LGS $Ax = b$ mit $x, b \in \mathbb{R}^n$ und der regulären Matrix $A \in \mathbb{R}^{n,n}$ der Form

$$\begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.217 \\ 0.254 \end{pmatrix}.$$

Um einfach festzustellen, ob ein Vektor x Lösung des Systems ist, prüft man, ob das Residuum $r = b - Ax$ (man beachte den Vorzeichenwechsel im Vergleich zur bisherigen Notation $Ax - b$) einen Nullvektor liefert. Nehmen wir also zwei Kandidaten für die Lösung und zwar

$$\begin{aligned} \bar{x} &= (0.341,\ -0.087)^T, \\ \hat{x} &= (0.999,\ -1.001)^T, \end{aligned}$$

und berechnen dafür das Residuum. Es gilt entsprechend

$$\begin{aligned} \bar{r} &= (0.000001,\ 0)^T, \\ \hat{r} &= (0.001343,\ 0.001572)^T. \end{aligned}$$

Die "Güte" des Fehlers könnte den Betrachter dazu verleiten, $\bar{x}$ als den besseren Vorschlag zu akzeptieren. Das ist jedoch ein Trugschluss bei Kenntnis der exakten Lösung $x^* = (1, -1)^T$. Der Grund ist, dass die Matrix A eine schlechte Kondition hat. Kennzeichen dafür sind unter anderem:

- Die Matrix A ist fast singulär.
- Die Determinante von A ist nahe Null, denn $\det(A) = 10^{-6}$.
- Wenn die Matrix A Elemente der Größenordnung $\mathcal{O}(1)$ besitzt, dann hat die inverse Matrix A^{-1} betragsmäßig große Elemente, hier

$$A^{-1} = \begin{pmatrix} 659000 & -563000 \\ -913000 & 780000 \end{pmatrix}, \quad \det(A^{-1}) = 10^6.$$

– Das Spektrum der Eigenwerte von A ist sehr "breit". Es liegen Größenordnungen zwischen dem betragsmäßig kleinsten ($\neq 0$) und größten Eigenwert, denn sie betragen hier

$$\lambda_1 = 0.000\,000\,694\,927\,368,$$
$$\lambda_2 = 1.438\,999\,305\,072\,632.$$

– Eine einfache geometrische Interpretation ist die Charakterisierung der Lösung als Schnittpunkt zweier Geraden, die sich hier in einem extrem spitzen Winkel schneiden.

Mit der Zeilensummennorm $\|A\|_\infty$ beträgt die Kondition

$$\mathrm{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty \approx 2.5 \cdot 10^6.$$

Der Exponent $t = 6$ in der Näherung des Konditionswerts charakterisiert im Groben die Reduzierung der gültigen Mantissenstellen der in Rechnungen benutzten GPA.

Rechnungen in Maple

```
> restart:  with(linalg):
```

Definition des LGS mit Matrix A (Float-Zahlen bzw. symbolisch)

```
> Digits:=10:
  A:=matrix(2,2,[0.780,0.563,0.913,0.659]);
  AA:=matrix(2,2,(i,j)->convert(A[i,j],rational));
  rank(A);
  det(A);
  norm(A);
  inverse(A);
  inverse(AA);
  b:=vector([0.217,0.254]);
  xs:=vector([1,-1]);
 `Ax*-b`=evalm(A&*xs-b);
```

$$A := \begin{bmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{bmatrix}$$

$$AA := \begin{bmatrix} \frac{39}{50} & \frac{563}{1000} \\ \frac{913}{1000} & \frac{659}{1000} \end{bmatrix}$$

$$2$$

$$0.1\,10^{-5}$$

$$1.572$$

$$\begin{bmatrix} 659000.0000 & -563000.0000 \\ -913000.0000 & 780000.0000 \end{bmatrix}$$

$$\begin{bmatrix} 659000 & -563000 \\ -913000 & 780000 \end{bmatrix}$$

$$b := [0.217,\ 0.254]$$

$$xs := [1, -1]$$
$$Ax^* - b = [0., 0.]$$

Rechnung mit schwacher GPA

```
> Digits:=5;
  evalm(A);
  rank(A);
  det(A);
  norm(A);
  inverse(A);
```

$$Digits := 5$$
$$\begin{bmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{bmatrix}$$
$$1$$
$$0.$$
$$1.572$$

```
Error, (in inverse) singular matrix
```

Test von zwei Lösungskandidaten und Berechnung der Residuen dazu

```
> Digits:=10;
  xq := vector([0.341,-0.087]);
  xd := vector([0.999,-1.001]);
  multiply(A,xq);
  rq := evalm(b-multiply(A,xq));
  rd := evalm(b-A&*xd);
```

$$Digits := 10$$
$$xq := [0.341, -0.087]$$
$$xd := [0.999, -1.001]$$
$$[0.216999, 0.254000]$$
$$rq := [0.1\,10^{-5}, 0.]$$
$$rd := [0.001343, 0.001572]$$

Lösung des LGS mit GPA bei unterschiedlicher Mantissenlänge

```
> Digits:=16;
  erg:=linsolve(A,b);
  erg[1]; erg[2];
  i:='i':
  printf(‘Digits         x[1]                      x[2]\\n‘):
  for i from 5 to 16 do
   Digits:=i:
   erg:=linsolve(A,b):
   if rank(A)<2 then
    printf(‘%2d       Matrix A singulaer\n‘,i)
   else
    printf(‘%2d       %.16e  %.16e\n‘,i,erg[1],erg[2])
   end if:
  end do:
```

$$Digits := 16$$
$$erg := [1.000000000000000,\ -1.000000000000000]$$
$$1.000000000000000$$
$$-1.000000000000000$$

```
Digits        x[1]                        x[2]
 5       Matrix A singulaer
 6       Matrix A singulaer
 7       Matrix A singulaer
 8       9.7422161000000000e-01  -9.6428571000000000e-01
 9       1.0006597800000000e+00  -1.0009140800000000e+00
10       9.9993410670000000e-01  -9.9990870920000000e-01
11       1.0000131801000000e+00  -1.0000182602000000e+00
12       9.9999802300600000e-01  -9.9999726100400000e-01
13       9.9999986820020000e-01  -9.9999981740000000e-01
14       9.9999998023004000e-01  -9.9999997261000000e-01
15       9.9999999934100200e-01  -9.9999999908700000e-01
16       1.0000000000000000e+00  -1.0000000000000000e+00
```

Beispiel 2.39 Typische Beispiele für Genauigkeitsuntersuchungen unter Beachtung der Kondition sind Tridiagonalsysteme, wie sie in Kap. 1.1 erläutert worden sind. Wir betrachten also das LGS $Ax = b$ mit

$$A(n,n) \;=\; \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \;=\; \text{tridiag}(-1, 2, -1).$$

Die reellen positiven EW der spd und irreduzibel diagonaldominanten Matrix A sind

$$\lambda_i = 2[1 - \cos(i\pi/(n+1))] = 4\sin^2(i\pi/(2(n+1))), \quad i = 1, 2, ..., n,$$

die EV

$$v^{(i)} = (v_1^{(i)}, v_2^{(i)}, ..., v_n^{(i)})^T \quad \text{mit} \quad v_j^{(i)} = \sin(ij\pi/(n+1)).$$

Mit $h = 1/(n+1)$, $\cos(x) = 1 - 2\sin^2(x/2)$, gelten die Beziehungen

$$\lambda_i = 4\sin^2(i\pi h/2),$$

$$v^{(i)}, \quad v_j^{(i)} = \sin(ij\pi h),$$

$$0 < 2(1 - \cos(\pi h)) = 4\sin^2(\pi h/2) = \lambda_{min} = \lambda_1 < \lambda_2 < ... < \lambda_n = \lambda_{max} =$$

$$= 4\sin^2(n\pi h/2) = 4[1 - \sin^2(\pi h/2)] = 4\cos^2(\pi h/2) = 4 - \lambda_{min} < 4,$$

und die Abschätzungen

$$\lambda_1 \approx \pi^2 h^2, \qquad \lim_{h\to 0} \lambda_1 = 0,$$
$$\lambda_n \approx 4 - \pi^2 h^2, \quad \lim_{h\to 0} \lambda_n = 4,$$
$$\lambda_{\frac{n+1}{2}} = 2, \quad \text{falls } n \text{ ungerade},$$
$$\sigma(A) \in (0, 4).$$

Leider ist die Kondition der Matrix A sehr schlecht, d. h. wir haben mit der Spektralnorm

$$\begin{aligned} \|A\|_2 &= \sqrt{\max_{i=1(1)n} \mu_i}, \quad 0 \le \mu_i \in \sigma(A^T A), \\ &= \sqrt{\rho(A^T A)}, \\ &\qquad \sigma(A^T A) = \sigma(A A^T) = \{\mu_i(A^T A),\ i = 1, 2, ..., n\} \ \text{ Spektrum}, \end{aligned}$$

wegen $A = A^T > 0$ die spektrale Kondition

$$\begin{aligned} \kappa(A) &= \operatorname{cond}_2(A) = \|A\|_2 \, \|A^{-1}\|_2 \\ &= \frac{\lambda_{max}}{\lambda_{min}} = \frac{4 - \lambda_{min}}{\lambda_{min}} = \frac{4}{\lambda_{min}} - 1 \\ &\approx \frac{4}{4 \sin^2(\pi h/2)} \approx \frac{1}{(\pi h/2)^2} \\ &\approx \frac{4n^2}{\pi^2} \gg 1 \quad \text{für} \quad n \gg 1. \end{aligned}$$

Um eine akzeptable Näherungslösung $\tilde{x}$ zur exakten Lösung x zu erhalten, ist eine starke GPA notwendig.
Die Fehlerakkumulation beim GA mit $A(n, n)$ und t Dualstellen der Mantisse des GPF führt bei $n \to \infty$ zum absoluten Fehler

$$\|\Delta x\| = \|\tilde{x} - x\| = 2^{-t} \operatorname{cond}(A)\, K(n),$$

wobei für den verkürzten GA ("chase method", "metod progonka") ohne Pivotisierung bei $A = \text{tridiag}()$ und diagonaldominant oder $A = A^T > 0$ die Beziehung

$$K(n) = \mathcal{O}(\sqrt{n})$$

gilt.
Für $t = 64$ (*extended*-Format), $t = 53$ (*double*-Format) oder $t = 40$ (*real*-Format) kann man die gültigen Dezimalstellen der Näherungslösung ermitteln.

Literaturverzeichnis

[1] Beresin, I. S. und N. P. Shidkow: *Numerische Methoden*. Bd. 1,2. DVW Berlin 1970, 1971.

[2] Bodewig, E.: *Matrix calculus*. New York 1959.

[3] Collatz, L.: *Funktionalanalysis und numerische Mathematik*. Springer-Verlag Berlin 1964.

[4] Demidovich, B. P., I. A. Maron und E. S. Schuwalowa: *Numerische Methoden der Analysis*. Math. für Naturwiss. und Technik, Bd. 14. DVW Berlin 1968.

[5] Deuflhard, P. und H. Hohmann: *Numerische Mathematik*. Walter de Gruyter Berlin 1991.

[6] Deuflhard, P. und H. Hohmann: *Numerische Mathematik*. 1: Eine algorithmisch orientierte Einführung. 3. überarbeitete und erweiterte Auflage, Lehrbuch. Walter de Gruyter Berlin 2002.

[7] Dietel, J.: *Formelsammlung zu Numerischen Mathematik mit Turbo Pascal-Programmen* (TPNUM). Rechenzentrum der RWTH Aachen 1993.

[8] Donner, K.: *Skalierung von Matrizen und numerische Stabilität der Gauß-Elimination*. Preprint Universität Passau, MIP-9514 September 1995.

[9] Faddejew, D. K. und W. N. Faddejewa: *Numerische Methoden der linearen Algebra*. Math. für Naturwiss. und Technik, Bd. 10. DVW Berlin 1973.

[10] Faires, J. D. und R. L. Burden: *Numerische Methoden. Näherungsverfahren und ihre praktische Anwendung*. Spektrum Akademischer Verlag Heidelberg 1994.

[11] Gantmacher, F.R.: *Teorija matriz*. Moskwa 1954.

[12] Golub, G. H. und C. van Loan: *Matrix Computations*. The John Hopkins University Press Baltimore 1989.

[13] Hackbusch, W.: *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Leitfäden der angewandten Mathematik und Mechanik Band 69. B. G. Teubner Stuttgart 1991, 1993.

[14] Hanke-Bourgeois, M.: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Mathematische Leitfäden. B. G. Teubner GmbH, Stuttgart 2002.

[15] Hämmerlin G. und K.-H. Hoffmann: *Numerische Mathematik*. Grundwissen Mathematik 7. Springer-Verlag Berlin 1991, 1994.

[16] Hermann, M.: *Numerische Mathematik*. R. Oldenbourg Verlag München 2001.

[17] Huckle, T. unds. Schneider: *Numerik für Informatiker*. Springer-Verlag 2002.

[18] Isaacson, E. und H. B. Keller: *Analyse numerischer Verfahren*. Edition Leipzig 1972.

[19] Kiełbasiński, A. und H. Schwetlick: *Numerische lineare Algebra*. DVW Berlin 1988.

[20] Kose, K., R. Schröder und K. Wieliczek: *Numerik sehen und verstehen*. Ein kombiniertes Lehr- und Arbeitsbuch mit Visualisierungssoftware. Vieweg Braunschweig 1992.

[21] Lax, P. D.: *Linear Algebra*. Pure and Applied Mathematics. John Wiley & Sons New York 1997.

[22] Locher, F.: *Numerische Mathematik - für Informatiker*. Springer-Verlag Berlin 1992.

[23] Luetkepohl, H.: *Handbook of matrices*. John Wiley & Sons Chichester 1996.

[24] Maess, G.: *Vorlesungen über Numerische Mathematik I, II*. Akademie-Verlag Berlin 1984, 1988. Verfahren. Vieweg Braunschweig 1999.

[25] Meis, Th. und U. Marcowitz: *Numerische Behandlung partieller Differentialgleichungen*. Springer-Verlag Berlin 1978.

[26] Meister, A.: *Numerik linearer Gleichungssysteme*. Ein Einführung in moderne

[27] Mohr, R.: *Numerische Methoden in der Technik*. Ein Lehrbuch mit Matlab-Routinen. Vieweg Wiesbaden 1998.

[28] Neundorf, W.: *Numerische Mathematik - Vorlesungen, Übungen, Algorithmen und Programme*. Shaker Verlag Aachen 2002.

[29] Oevel, W.: *Einführung in die Numerische Mathematik*. Spektrum Akademischer Verlag Heidelberg 1996.

[30] Opfer, G.: *Numerische Mathematik für Anfänger*. Vieweg Studium Grundkurs Mathematik Wiesbaden 1993, 3. überarbeitete und erw. Auflage 2001.

[31] Paulin, G. und E. Griepentrog: *Numerische Verfahren der Programmiertechnik*. Verlag Technik Berlin 1975.

[32] Plato, R.: *Numerische Mathematik kompakt*. Grundlagenwissen für Studium und Praxis. Vieweg Wiesbaden 2000.

[33] Quarteroni, A., Sacco, R. und F. Saleri: *Numerische Mathematik 1,2*. Springer-Verlag 2002.

[34] Ralston, A.: *A First Course in Numerical Analysis*. McGraw-Hill New York 1965.

[35] Reimer, M.: *Grundlagen der Numerischen Mathematik*. I und II. AULA-Verlag Wiesbaden 1980, 1982.

[36] Rice, J. R.: *Numerical Methods, Software and Analysis*. 2nd Edition. Academic Press Inc. Boston 1993.

[37] Roos, H.-G. und H. Schwetlick: *Numerische Mathematik*. Das Grundwissen für jedermann. B. G. Teubner Stuttgart 1999.

[38] Rutishauser, H.: *Vorlesungen über Numerische Mathematik*. Bd. 1,2. Birkhäuser Verlag Basel 1976.

[39] Samarskij, A. A.: *Theorie der Differenzenverfahren*. Akademische VG Geest & Portig K.-G. Leipzig 1984.

[40] Schaback, R. und H. Werner: *Numerische Mathematik*. Springer-Verlag Berlin 1993.

[41] Schuppar, B.: *Elementare Numerische Mathematik*. Eine problemorientierte Einführung für Lehrer und Studierende. Vieweg Wiesbaden 1998.

[42] Schwarz, H. R.: *Numerische Mathematik*. B. G. Teubner Stuttgart 1988, 1997.

[43] Schwarz, H. R., H. Rutishauser und E. Stiefel: *Numerik symmetrischer Matrizen*. B. G. Teubner Stuttgart 1972.

[44] Schwetlick, H. und H. Kretzschmar: *Numerische Verfahren für Naturwissenschaftler und Ingenieure*. Fachbuchverlag Leipzig 1991.

[45] Späth, H.: *Numerik*. Vieweg Wiesbaden 1994.

[46] Stoer, J.: *Numerische Mathematik I*. Heidelberger Taschenbücher 105. Springer-Verlag Berlin 1993.

[47] Stoer, J. und R. Bulirsch: *Einführung in die Numerische Mathematik II*. Heidelberger Taschenbücher 114. Springer-Verlag Berlin 1990.

[48] Stoer, J. und R. Bulirsch: *Numerical mathematics 2*. An Introduction - under consideration of lectures by F. L. Bauer. 4. neu bearbeitete und erweiterte Auflage. Springer-Verlag Berlin 2000.

[49] Überhuber, C.: *Computer-Numerik 1,2*. Springer-Verlag Berlin 1995.

[50] Varga, R.S.: *Matrix iterative analysis*. Prentice Hall, Englewood Cliffs, N. J., 1962.

[51] Weller, F.: *Numerische Mathematik für Ingenieure und Naturwissenschaftler*. Eine Einführung für Studium und Praxis. Vieweg Braunschweig 1996.

[52] Werner, J.: *Numerische Mathematik 1, 2*. Vieweg Studium Aufbaukurs Mathematik 1992.

[53] Wilkinson, J. H. und C. Reinsch: *Linear Algebra*. Handbook for automatic computation, Vol. II. Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Bd. 186. Springer-Verlag Berlin 1971.

[54] Zlatev, Z.: *Computational Methods for General Sparse Matrices*. Math. and Its Appl. Vol.65. Kluwer Academic Publishers London 1991.

[55] Zurmühl, R.: *Praktische Mathematik für Ingenieure und Physiker*. Springer-Verlag Berlin 1965.

[56] Zurmühl, R. und S. Falk: *Matrizen und ihre Anwendungen*. Teil 2, Numerische Methoden. Springer-Verlag Berlin 1984.

Symbolverzeichnis

$\mathbb{N} = \{0, 1, ...\}$	Menge der natürlichen Zahlen
$\mathbb{Z} = \{..., -1, 0, 1, ...\}$	Menge der ganzen Zahlen
$\mathbb{Q}$	Körper der rationalen Zahlen
$\mathbb{R}$	Körper der reellen Zahlen
$\mathbb{R}_+$	Menge der positiven reellen Zahlen
$\mathbb{R}^n$	n-dimensionaler Raum der Vektoren $x = (x_1, ..., x_n)^T$
$\mathbb{C}$	Körper der komplexen Zahlen
$\mathbb{C}^n$	n-dimensionaler Raum der komplexen Vektoren
$\perp,\ \not\perp$	Orthogonalität von Vektoren aus $\mathbb{R}^n$ bzw. nicht orth.
$i, \imath$	imaginäre Einheit $\sqrt{-1}$
I, $(a, b),\ [a, b]$	offenes bzw. abgeschlossenes Intervall auf $\mathbb{R}$
$\mathbb{R}^{m,n},\ \mathbb{R}^{m\times n}$	Menge der reellen Rechteckmatrizen mit m Zeilen und n Spalten
$\mathcal{M},\ \mathcal{L},\ \mathcal{X}$	Menge, Raum, linearer (normierter) Raum
$\lvert\mathcal{M}\rvert$	Mächtigkeit der Menge $\mathcal{M}$
$\mathcal{M}^\perp$	orthogonales Komplement der Menge $\mathcal{M}$
$\dim(\mathcal{M})$	Dimension des Raums $\mathcal{M}$
$\mathcal{C}[a, b] = \mathcal{C}^0[a, b],\ \mathcal{C}([a, b], \mathbb{R})$	Raum der reellen stetigen Funktionen auf $[a, b]$
$\mathcal{C}^k(\mathbb{R}),\ \mathcal{C}^k(\mathbb{R}, \mathbb{R}),\ \mathcal{C}^k(\mathrm{I})$	Raum der reellen k-mal stetig differenzierbaren Funktionen über $\mathbb{R}$ bzw. I
$I,\ I(n, n),\ I_n$	Einheitsmatrix
$A = (a_{ij}),\ A = (a_{i,j})_{i,j=1}^n$	Matrix mit ihren Elementen
A^T	Transponierte der Matrix A
A^{-1}	Inverse der Matrix A
$\bar{A}^T$	zu A konjugiert komplexe und transponierte Matrix
A^*	zu A adjungierte Matrix, $A^* = A^H = \bar{A}^T$
A^+	Pseudoinverse
$\det(A)$	Determinante der Matrix
$\dim(A)$	Dimension der (quadratischen) Matrix
$\mathrm{rang}(A),\ \mathrm{rank}(A)$	Rang der Matrix
$\mathrm{spur}(A),\ \mathrm{trace}(A)$	Spur der Matrix

$\mathrm{diag}(A)$	Matrix aus den Diagonalelementen von A
$\mathrm{diag}(d_1, d_2, ..., d_n)$	Diagonalmatrix mit n Zeilen und Spalten sowie Diagonalelementen d_i
$\mathrm{tridiag}(a_i, b_i, c_i)$	Tridiagonalmatrix $A = (a_{ij})$ mit n Zeilen und Spalten sowie Elementen $a_{ii} = b_i,\ a_{i,i+1} = c_i,\ a_{i,i-1} = a_i$
$(\cdot, \cdot)$	Skalarprodukt in $\mathbb{R}^n,\ \mathbb{C}^n,\ L_{2,\omega}$ oder l_2
$(x, x)_A = (Ax, x)$	spezielle Bilinearform in $\mathbb{R}^n$ mit $A = A^T > 0$
$\|\cdot\|$	Norm
$\|\cdot\|_s$	spezielle Norm
$\mathrm{cond}(A),\ \kappa(A)$	relative Konditionszahl der Matrix
$\mathrm{cond}_s(A)$	normabhängige relative Konditionszahl der Matrix
$\mathrm{acond}(A)$	absolute Konditionszahl der Matrix
$\mathrm{hcond}(A)$	Konditionszahl der Matrix nach Hadamard
$\mathrm{fcond}(A)$	Konditionszahl der Matrix nach Forsythe/Moler
$R(x) = \frac{x^T A x}{x^T x},\ x \neq 0$	Rayleigh-Quotient
$Ax = \lambda x,\ x \neq 0$	Eigenwertproblem für Matrix A
$\lambda(A),\ \lambda_i(A),\ \mu(A)$	Eigenwert der Matrix A
$\sigma(A)$	Spektrum der Matrix A, $\sigma(A) = \{\lambda(A)\}$
$\rho(A)$	Spektralradius der Matrix A, $\rho(A) = \max \lvert\lambda(A)\rvert$
$R_m(H) = -\frac{1}{m} \ln(\|H^m\|)$	mittlere Konvergenzrate
$R(H) = -\ln(\rho(H))$	(asymptotische) Konvergenzrate
$\Re(z),\ \Im(z)$	Real- bzw. Imaginärteil von z
$\Omega,\ \overline{\Omega}, \mathrm{int}(\Omega)$	Gebiet des $\mathbb{R}^n$, sein Abschluss und Inneres
$\partial\Omega = \overline{\Omega} \backslash \mathrm{int}(\Omega)$	Rand des Gebiets
$f'(x), f''(x), f^{(k)}(x), \dot{f}(x)$	klassische Ableitungen von $f(x)$
$f'(x) = \mathcal{J}(x) = (\frac{\partial f_i(x)}{\partial x_j}),\ \mathcal{G}(x)$	Jacobi-Matrizen der Vektorfunktionen $f, g : \mathbb{R}^n \to \mathbb{R}^n$
$\mathrm{grad}\ Q(x) = \nabla Q(x) = (\frac{\partial Q(x)}{\partial x_i})$	Gradient des Funktionals $Q : \mathbb{R}^n \to \mathbb{R}$
$\nabla^2 Q(x) = (\frac{\partial^2 Q(x)}{\partial x_i \partial x_j})$	Hesse-Matrix des Funktionals $Q : \mathbb{R}^n \to \mathbb{R}$
Δ	Laplace-Operator
div	Divergenz
$\mathrm{sign}(x),\ \mathrm{sign}(f(x))$	Vorzeichen von $x \in \mathbb{R}$ bzw. des Funktionswerts $f(x)$
$\mathrm{sign1}(x)$	Vorzeichen von $x \in \mathbb{R}$ für $x \neq 0$, ansonsten 1
$\mathrm{span}\{f_k\}$	Raum der Funktionen aufgespannt durch f_k, lineare Hülle

$\mathcal{X} = [\mathrm{X}] = \mathrm{span}\{x_1, ..., x_n\}$	Vektorraum aufgespannt durch $x_k \in \mathbb{R}^n$
$\mathcal{K}_k = \mathcal{K}_k(A, x) = \mathrm{span}\{x, Ax, ..., A^{k-1}x\}$	Krylov-Unterraum
$LK(x, y),\ LK(z_k)$	Linearkombination von Vektoren x, y, z_k
$\mathcal{N}(L) = \{x \mid Lx = 0\}$	Nullraum des Operators L
$T(n),\ K(n)$	Komplexität eines Problems der Dimension n
$[\,\cdot\,]$	Gauß-Klammer, $[x]$ ist die größte ganze Zahl $\leq x$
$n! = 1 \cdot 2 \cdot ... \cdot n$	Fakultät
$(2n)!!$	$(2n)!! = 2 \cdot 4 \cdot ... \cdot 2n$
$(2n-1)!!$	$(2n-1)!! = 1 \cdot 3 \cdot ... \cdot (2n-1)$
$\binom{n}{k} = \frac{n(n-1)\cdot ... \cdot (n-k+1)}{k!}$	Binomialkoeffizient
$\mathcal{P}_n$	$(n+1)$-dimensionaler Raum aller Polynome vom Grad $\leq n$ über dem Körper $\mathbb{R}$
$p_n(x), q_m(x), r_k(t), s(t), ...$	Polynome aus $\mathcal{P}_n$
$\mathrm{Grad}(p_n)$	Grad des Polynoms $p_n(x)$
$\mathrm{int}(x_0, x_1, ..., x_n)$	offenes Intervall aufgespannt durch die Punkte $x_i \in \mathbb{R}$
$\{x_0, x_1, ..., x_n\}$	Stützstellenfolge, evtl. x_i paarweise verschieden
$R = \{(x_i, y_i) \mid a \leq x_i \leq b, \quad i = 0, 1, ..., n\}$	Referenz, evtl. x_i paarweise verschieden
$R_0 = \{(x_i, y_i) \mid a \leq x_0 < ... < x_n \leq b\}$	Referenz, x_i geordnet
$\mathcal{S}_m(R)$	Raum der Splines vom Grad $\leq m$ zur Referenz R
$\Phi = (\varphi_j(x_i))_{i=0,\ j=0}^{N,\ n}$	Haarsche Matrix des Funktionensystems $\{\varphi_j\}$ mit Referenz R
$\omega = e^{\imath 2\pi/n}$	komplexe Wurzel
$\liminf_{n\to\infty} x_n$	$\underline{\lim}\ x_n$, limes inferior, kleinster Häufungspunkt der Folge $\{x_n\}$
$\limsup_{n\to\infty} x_n$	$\overline{\lim}\ x_n$, limes superior, größter Häufungspunkt der Folge $\{x_n\}$
δ_{ij}	Kronecker-Symbol, $\delta_{ij} = 1$, falls $i = j$, sonst Null
$\mathcal{O}(\cdot),\ o(\cdot)$	Landau-Symbole
$\emptyset$	leere Menge
$\gg,\ \ll$	im Vergleich viel größer bzw. viel kleiner
$\square$	Ende eines Beweises
$\rightarrow,\ \Rightarrow$	daraus folgt
$\Rightarrow,\ \Leftarrow$	Hin- bzw. Rückrichtung im Beweis
$\leftrightarrow$	Tausch von Größen

Akronyme und Abkürzungen

o.B.d.A.	ohne Beschränkung der Annahme
gdw.	genau dann wenn
spd	symmetrische, positiv definite Matrix, $A = A^T > 0$
GPZ	Gleitpunktzahl
GPF	Gleitpunktformat
GPA	Gleitpunktarithmetik
DGL	Differentialgleichung
LGS	lineares Gleichungssystem
GA	Gauß-Algorithmus, Gauß-Elimination
VGA	verketteter Gauß-Algorithmus
PE	Pivotelement
ZV	Zeilenvertauschung
IV	Iterationsverfahren
GSV, JA	Gesamtschrittverfahren, Jacobi-IV
ESV, GS	Einzelschrittverfahren, Gauß-Seidel-IV
NNE	Nichtnullelement
RWA	Randwertaufgabe, Zweipunktrandwertaufgabe
DS	Differenzenschema
FDM	finite Differenzenmethode
FEM	Finite-Elemente-Methode
CAS	Computeralgebrasystem
TP	Turbo Pascal
EWP	Eigenwertproblem
EW	Eigenwert
EV	Eigenvektor
OGS	Orthogonalsystem von Vektoren/Funktionen
ONS	Orthonormalsystem
FT	Fourier-Transformation
GGT	größter gemeinsamer Teiler ganzer Zahlen
Matlab	MATrix LABoratory
AA	Approximationsaufgabe
IA	Interpolationsaufgabe

Index

Anschrift:

Dr. rer. nat. habil. Werner Neundorf
Technische Universität Ilmenau, Institut für Mathematik
PF 10 05 65
D - 98684 Ilmenau

E-mail : werner.neundorf@tu-ilmenau.de
Homepage : http://www.mathematik.tu-ilmenau.de/∼neundorf/index_de.html